TYPE I ERROR AND POWER OF THE MEAN AND COVARIANCE STRUCTURE CONFIRMATORY FACTOR ANALYSIS FOR DIFFERENTIAL ITEM FUNCTIONING DETECTION: METHODOLOGICAL ISSUES AND RESOLUTIONS

BY

Jaehoon Lee

Submitted to the graduate degree program in Psychology
and the Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy.

_____	_____
Todd D. Little (Co-Chair)	Kristopher J. Preacher (Co-Chair)
_____	_____
John Colombo (Committee Member)	Neal Kingston (Committee Member)
_____	
William P. Skorupski (Committee Member)	

Date defended: April 22, 2009

The Dissertation Committee for Jaehoon Lee certifies

that this is the approved version of the following dissertation:

TYPE I ERROR AND POWER OF THE MEAN AND COVARIANCE STRUCTURE CONFIRMATORY FACTOR ANALYSIS FOR DIFFERENTIAL ITEM FUNCTIONING DETECTION: METHODOLOGICAL ISSUES AND RESOLUTIONS

_____	_____
Todd D. Little (Co-Chair)	Kristopher J. Preacher (Co-Chair)
_____	_____
John Colombo (Committee Member)	Neal Kingston (Committee Member)
_____	
William P. Skorupski (Committee Member)	

Date approved: April 22, 2009

ABSTRACT

Recently, mean and covariance structure (MACS) confirmatory factor analysis (CFA) has been widely used to detect items with differential item functioning (DIF). Although how we define the scale does not impact overall model fit or tests for whether or not a given level of measurement equivalence holds, different scaling methods can lead to different conclusions when a researcher locates DIF in a scale. This dissertation evaluates the MACS analysis for DIF detection by means of a Monte Carlo simulation. The simulation results indicate that three statistically equivalent scaling methods provide different outcomes of DIF analysis. In addition, Bonferroni-correction improves the accuracy of the analysis, notably when a scale (or an anchor) is contaminated by DIF. Based on the previous and current simulation studies, this dissertation offers practical guidance for researchers who attempt to evaluate measurement equivalence using CFA.

Keyword: DIF, MACS, scaling, biased anchor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

There are many people who I would like to thank for their help on this dissertation. First of all, I gratefully acknowledge my mentor and supervisor, Dr. Todd D. Little, for his great support, guidance, and enthusiasm during my time at the University of Kansas. It was an honor that he gladly accepted to be my academic advisor when my first advisor, Dr. Susan E. Embretson, accepted a position at the Georgia Institute of Technology. His assistance has been invaluable and working with him has been a rich and rewarding experience. I am sure that I could never express the depth of my appreciation as he has had a tremendous influence on my life.

Additionally, I would like to acknowledge Dr. Kristopher J. Preacher for giving his statistical expertise and insightful suggestions throughout all stages of this dissertation. He greatly assisted me programming a simulation study. Without him, I would not be able to complete this dissertation. Also, I would like to thank my committee members, Dr. John Colombo, Dr. Neal Kingston, and Dr. William P. Skorupski for their thoughtful remarks on various aspects of this dissertation.

On a personal note, I would like to thank my family for their unwavering support and confidence as I pursued my graduate degrees in the United States. They always cared for my comfort and happiness. Additional thanks are due to my friends in Lawrence who have kept me laughing and enjoying life. I consider myself very lucky to know them.

CHAPTER I: INTRODUCTION

This chapter briefly introduces the concept of measurement equivalence. This is followed by a description of potential problems of using confirmatory factor analysis (CFA) to detect items with differential item functioning (DIF). At the end of the chapter, the purpose and structure of this dissertation are presented.

1. Measurement Equivalence

Measurement equivalence can be thought of as characteristics of an item or items that yield a test[1] of the same attribute under different conditions (Horn & McArdle, 1992). These conditions include different groups, administrations, and media (e.g., paper-based test versus computer-based test). With more than one group, a scale is said to have measurement equivalence when examinees with identical scores on the underlying (latent) construct but different group membership have the same observed or expected raw scores at the item level, at the scale level, or at both (Drasgow & Kanfer, 1985).

Applied psychologists have highlighted the importance of measurement equivalence as a prerequisite for meaningful group comparisons (e.g., Drasgow, 1984; Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993; Vandenberg, 2002; Vandenberg & Lance, 2000). To the extent that a set of items or a scale does not function equivalently across groups, any interpretation of group differences is necessarily open to question (Byrne & Stewart, 2006; Raju et al., 2002). For example, under ideal circumstances, observed mean differences represent true mean differences

[1] The terms test and scale are used synonymously in this dissertation.

across groups. However, when measurement equivalence is not defendable, they may represent true mean differences or differences in the psychometric relation that connects the observed responses to the latent construct. As such, a lack of measurement equivalence, or equivalently differential functioning, constitutes a potential threat to the validity of a scale. Accordingly, both APA (American Psychological Association) and ITC (International Test Commission) standards have emphasized evaluation of DIF for fair use of a scale (AREA, APA, NCME, 1999).

2. Potential Problems of Using CFA for DIF Detection

Recently, CFA has enjoyed increasing attention from DIF researchers. For example, a particular form of mean and covariance structure (MACS; Sörbom, 1974) CFA has been widely used to test measurement equivalence at item level (e.g., Byrne, 1998; Chan, 2000; Everson, Millsap, & Rodriguez, 1991; Ferrando, 1996; González-Romá, Tomás, Ferreres, & Hernández, 2005; Wasti, Bergman, Glomb, & Drasgow, 2000). Supporting the utility of the CFA approach, previous simulation studies have shown that the MACS analysis works fairly well for testing DIF under various conditions (e.g., Finch, 2005; González-Romá, Hernández, & Gómez-Benito, 2006; Hernández & González-Romá, 2003; Meade & Lautenschlager, 2004; Navas-Arai & Gómez-Benito, 2002; Oort, 1998; Stark, Chernyshenko, & Drasgow, 2006; Wanichtanom, 2001).

In any CFA model, the scale for the item or latent construct parameters needs to be identified in order to yield unique estimates of the parameters (Jöreskog & Sörbom, 1989). In multiple-group cases, the scaling has been achieved conventionally

by choosing an anchor item whose parameters are constrained to be equal across groups. Unfortunately, such practice implies a tacit assumption of parameter invariance, even when the purpose of analysis is to evaluate measurement equivalence (Cheung & Rensvold, 1999). If the invariance assumption is not tenable, any analysis may not provide a proper solution (Bollen, 1989). This, in turn, may lead to inaccurate conclusions about DIF with regard to the other items being tested within the scale (Cheung & Rensvold, 1999; Millsap, 2005). Thus, researchers need to acknowledge the potential problems of choosing a conventional scaling method, especially when they conduct CFA for DIF analysis.

3. Purpose and Structure of the Dissertation

The purpose of this dissertation is to examine the statistical properties of the MACS analysis for DIF detection. More specifically, this dissertation evaluates the Type I error and power of this methodology under a variety of conditions. Based on the results from a simulation study, this dissertation proposes an analytic strategy that is robust to the misspecification problems as well as theoretically suitable for DIF analysis.

The structure of this dissertation is as follows. Chapter II briefly demonstrates basic concepts and terminologies that have been developed in the DIF literature. This chapter closes with a detailed description of CFA and its methodologies used for DIF detection. Chapter III presents a Monte Carlo study that assesses the Type I error and power of the MACS technique. This chapter includes design of the study, discussion of the manipulated variables, and procedures used to generate data and run the

analyses. The results of the simulation study are presented in Chapter IV. Chapter V discusses the simulation results in the context of educational and psychological assessment. In addition, limitations of the study are described and future directions for research are considered.

CHAPTER II: LITERATURE REVIEW

Among various techniques for testing measurement equivalence, those most commonly used are based on item response theory (IRT) or confirmatory factor analysis (CFA), or more broadly structural equation modeling (SEM) (Teresi, 2006). The formal similarity between IRT and CFA has been introduced repeatedly in the literature (e.g., Mellenbergh, 1994; Muthén, 1984, 1989, 1993; Muthén & Christoffersson, 1981; Muthén & Lehman, 1985; Takane & de Leeuw, 1987). That is, they are comparable in the sense that both postulate that an unobserved continuous variable (i.e., latent construct) influences a set of observed variables. More importantly, both provide a statistical framework within which between-group equality can be evaluated for the item parameters (see Chan, 2000; Ferrando, 1996; Raju et al., 2002; Reise et al., 1993). The most apparent difference is that IRT is often applied to observed responses on categorical variables, whereas CFA has traditionally been applied to observed covariances among continuous variables.

This chapter introduces the IRT framework, in which differential item functioning (DIF) was originally conceptualized, and then provides a detailed description of the CFA framework.

1. IRT Framework

1.1. IRT

IRT models (e.g., Lord & Novick, 1968) have been developed predominantly in education and psychology since the late 1960s. Focused on observed responses on the binary or ordinal items, these models define probabilistic, nonlinear relations of

the responses to latent constructs (i.e., ability, denoted by θ). The basic assumption in IRT is that a set of items assesses a single ability dimension (i.e., unidimensionality) but they are pairwise uncorrelated if ability level is held constant (i.e., local independence).

The ability score for a particular examinee is estimated based on his or her observed item responses, given a value of each item parameter. For each item, two types of item parameters are frequently estimated. The *attractiveness* or *b* parameter determines the horizontal position of the item trace line, called the *item characteristic curve* (ICC) or *item response function* (IRF), which depicts the probability of an item response along the ability level continuum. The *b* parameter is typically referred to as item *difficulty* in the cases of binary item (Lord, 1980); the higher the *b* parameter value, the more difficult it is to answer the item correctly. The *discrimination* or *a* parameter determines the slope of the ICC; the higher the *a* parameter value, the stronger the relationship is between ability level and response on the item. Therefore, an item with a substantial *a* parameter value can powerfully differentiate examinees with different ability scores.

The IRT models were originally developed for responses on the binary items, but in practice they also increasingly have been used for responses on polytomously ordered items. In fact, one-parameter and two-parameter models (Birnbaum, 1968; Hambleton, Swaminathan, & Rogers, 1991) for dichotomous response can be viewed as special cases of graded response model (GRM; Samejima, 1969) for polytomous responses. Thus, this dissertation uses GRM to illustrate how to assess measurement

equivalence in IRT. More details on polytomous models can be found in Bock (1972), Masters (1982), Muraki (1990), and Samejima (1969, 1972).

1.2. GRM

In GRM, the relationship between ability level and probability of endorsing any particular response option is graphically depicted by the *category response function* (CRF). The CRFs for each item are given by

$$P_{ik}(\theta_j) = \frac{e^{a_i(\theta_j - b_{ik+1})} - e^{a_i(\theta_j - b_{ik})}}{\left(1+e^{a_i(\theta_j - b_{ik})}\right)\left(1+e^{a_i(\theta_j - b_{ik-1})}\right)}, \quad (1)$$

where $P_{ik}(\theta_j)$ is a probability that an examinee j with a given value on θ will respond to an item i with category k. It should be noted that the exponential terms are replaced by 1 and 0 for the lowest and highest response options, respectively. Furthermore, the relationship between ability level and likelihood of choosing a progressively increasing response option is depicted by a series of boundary response functions (BRFs). The BRFs for each item are given by

$$P^*_{ik}(\theta_j) = \frac{e^{a_i(\theta_j - b_{ik})}}{1+e^{a_i(\theta_j - b_{ik})}}, \quad (2)$$

where $P^*_{ik}(\theta_j)$ is the probability that an examinee j with a given value on θ will respond to an item i at or above a response option k. The BRFs are simplified to the IRF for two-parameter logistic (2-PL) model in the case of binary items (Birnbaum, 1968). If $a = 1$, the BRFs become the IRF for one-parameter logistic (1-PL) model or Rasch model (Hambleton et al., 1991).

As observed in Equation 2, the BRFs depend on θ parameter as well as *b* and *a* parameters. For a particular item with *k* response options, there will be $k-1$ BRFs and the *a* parameter is constrained to be equal across BRFs. Consequently, each item is characterized by one discrimination parameter and several $(k-1)$ attractiveness parameters in IRT.

1.3. DIF

1.3.1. Definition

In IRT, a lack of measurement equivalence is referred to as differential functioning. When the differential functioning occurs at item level, it is called *differential item functioning* (DIF). Specifically, DIF represents between-group differences in the probability of an item response when ability scores are on a common scale (Mellenberg, 1994). The defining feature of DIF is that the ability scores are placed on a common scale or "statistically matched" across groups (Angoff, 1993; Camilli & Shepard, 1994). One method of statistical matching is to score all examinees using the same BRFs. However, the resultant ability estimates can be biased unless measurement equivalence has been established in advance.

An item is said to have measurement equivalence if the item parameters are identical across groups (Raju et al., 2002). That is, for group *g* (g = 1, …, *G*),

$$a_i^1 = a_i^2 = \ldots = a_i^G \text{ and}$$

$$b_{i1}^1 = b_{i1}^2 = \ldots = b_{i1}^G;\ b_{i2}^1 = b_{i2}^2 = \ldots = b_{i2}^G;\ \ldots;\ b_{ik}^1 = b_{ik}^2 = \ldots = b_{ik}^G. \quad (3)$$

When the item parameters are equal across groups, the CRFs and BRFs are also equal for these groups. Thus, it is possible to assess DIF either at the item parameter level

or at the IRF level; once a model has been selected for parameter estimation, DIF detection is performed by observing the BRFs or by directly comparing the item parameters. Either the invariant item parameters or the invariant BRFs suggest that true item responses will be identical for different examinees with equal ability scores (Raju et al., 2002).

1.3.2. Types of DIF

DIF can be either uniform or non-uniform depending on the item parameter that differs across groups. *Uniform* DIF is present when only the *b* parameters differ across groups. *Non-uniform* DIF exists when the *a* parameter differs across groups, regardless of whether or not the *b* parameters are different.

1.3.3. Sources of DIF

DIF, if present, is indicative of either item bias or item impact. *Item bias* occurs when the source of DIF is irrelevant to the construct being measured (Camilli & Shepard, 1994). For example, when different ethnic groups with equal ability scores exhibit different probabilities of an item response, item bias is said to occur. For item bias to be present, DIF must be apparent. Thus, DIF is a necessary, but not sufficient, condition for the item bias (Zumbo, 1999).

In contrast, *item impact* occurs when the source of DIF is a relevant characteristic of the construct being measured. In other words, item impact is evident when groups show different probabilities of an item response because they truly differ on the construct. In this case, the item parameter estimates of the test accurately reflect group differences in the construct.

1.4. IRT Methodologies for DIF Detection

Several IRT techniques have been proposed for DIF detection during the last two decades: Lord's (1980) chi-square, Raju's (1988) area method, and Thissen, Steinberg, and Wainer's (1988, 1993) IRT likelihood ratio (IRTLR). Only the IRTLR is illustrated here because it is most closely related to the CFA techniques (Cohen, Kim, & Baker, 1993; Cohen, Kim, & Wollack, 1996; Thissen, 1991; Thissen et al., 1988, 1993). Details for the other IRT techniques can be found in Lord (1980), Raju (1988, 1990), and Raju, van der Linden, and Fleer (1992).

The maximum likelihood (ML) parameter estimation algorithm results in a value of model fit. The IRTLR method assesses DIF by comparing the model fit of a pair of nested models. This technique starts with fitting the compact (simpler) model in which all item parameters are estimated with the constraint that they are equal across groups. Next, each item is tested, one at a time, for DIF. The augmented (complex) model relaxes the equality constraint for an item being tested. The latter model provides a value of the likelihood function, which is associated with estimating the parameters of the item being tested separately for each group. This value is compared to the value for the compact model by creating the likelihood ratio:

$$\frac{L_C}{L_{Ai}}. \quad (4)$$

Under the null hypothesis that the compact model holds in the population, –2 times the natural-log transformation of this ratio,

$$-2\ln\left(\frac{L_C}{L_{Ai}}\right) = -2(\ln L_C - \ln L_{Ai}), \quad (5)$$

yields a test statistic that approximately follows a chi-square distribution, with degrees of freedom equal to the difference in the number of estimated parameters between the two nested models. A significant chi-square value suggests that the compact model fits significantly worse than the augmented model. Equivalently, it is considered that the item parameters differ across groups; therefore, the item being tested exhibits DIF.

In a simulation study, Cohen et al. (1996) evaluated the IRTLR technique under various conditions. Although simulated responses were dichotomous rather than ordinal in nature, they found that, in general, this technique works reasonably well; Type I error rates fell within an expected range in most conditions.

In order to accurately estimate the IRT parameters, it is well known that a substantial number of items are required because the ability scores are predicted from the joint relationship with other items. In general, more than 30 items are recommended for stable parameter estimation in the literature (e.g., Seong, 1990; Stone, 1992).

2. CFA Framework

2.1. CFA

CFA models, which comprise the measurement component of the structural equation modeling (SEM), were developed in the 1970s mainly by sociologists and econometricians (see Jöreskog, 1971a, 1971b, 1973; McArdle & McDonald, 1984). The main objectives of CFA are to support hypothesis-driven data analysis as well as

to compare and refine theories on the basis of data, especially data obtained from non-experimental social research

Currently, the *mean and covariance structure* (MACS; Sörbom, 1974) model is ideally suited to evaluate measurement equivalence for several reasons (see Little, 1997). First, a hypothesized factor structure is fitted simultaneously in two or more groups. Second, it tests the between-group equality of all reliable measurement parameters. Third, it corrects for measurement error whereby estimates of the trait parameters are less biased. Finally, "strong" tests for measurement equivalence are tenable by evaluating the mean structure invariance of the observed responses. This dissertation uses the MACS model to demonstrate how to assess measurement equivalence in CFA. Although several CFA programs are now available, LISREL (Jöreskog & Sörbom, 1996) notation is used here for convenience.

2.2. MACS Model

The CFA models posit a linear, rather than a nonlinear, relation between observed responses and latent constructs (i.e., trait, denoted by ξ). In the MACS model, the observed response x_i to an item i (i = 1, …, p) is represented as a linear function of an intercept τ_i, latent trait variables ξ_i (j = 1, …, m), and a unique factor score δ_i. More specifically,

$$x_i = \tau_i + \lambda_{ij}\xi_i + \delta_i, \quad (6)$$

where the factor loading λ_{ij} defines the metric of measurement, as it represents the expected change in x_i per unit change in ξ_i. The intercept τ_i represents the expected value of x_i when $\xi_i = 0$. The unique factor score is further divided into two

components; an item-specific factor score and measurement error. The item-specific factor score represents systematic differences in an item response after influences of the trait variables have been eliminated. In contrast, the measurement error is typically conceptualized as random error. The unique factor score, or equivalently the sum of the item-specific factor score and the measurement error, is assumed to be normally distributed across observations.

When the same model holds in each group g (g = 1, …, G), Equation 6 is extended to

$$x^g = \tau^g + \Lambda^g \xi^g + \delta^g, \quad (7)$$

where x^g is a $p \times 1$ vector of observed responses (in group g), τ^g is a $p \times 1$ vector of intercepts, ξ^g is an $m \times 1$ vector of latent trait variables, Λ^g is a $p \times m$ matrix of factor loadings, and δ^g is a $p \times 1$ vector of unique factor scores.

In general, the MACS model assumes that (a) the unique factor scores are independent of the trait variables, (b) the unique factor scores are independent of each other, and (c) the expected unique factor scores are equal to zero. Under these assumptions, taking the expectation of Equation 7 yields the relation between the observed item means and the latent trait means:

$$\mu^g = \tau^g + \Lambda^g \kappa^g, \quad (8)$$

where μ^g is a $p \times 1$ vector of item means and κ^g is an $m \times 1$ vector of trait means for each group.

The covariance matrix of x variables is obtained in group g as

$$\Sigma^g = \Lambda^g \Phi^g \Lambda^{g'} + \Theta^g, \quad (9)$$

where Φ^g is an $m \times m$ covariance matrix of latent trait variables and Θ^g is a $p \times p$ matrix of unique factor score variances. This structural model is fitted to a sample covariance matrix S^g, yielding

$$S^g \approx \widehat{\Lambda^g}\widehat{\Phi^g}\widehat{\Lambda^{g'}} + \widehat{\Theta^g} = \widehat{\Sigma^g}, \quad (10)$$

where S^g is a $p \times p$ sample covariance matrix of x variables in group g and $\widehat{\Lambda^g}$, $\widehat{\Phi^g}$, and $\widehat{\Theta^g}$ matrices contain the estimates of population parameters. The sample covariance matrix is approximated by the CFA solution $\widehat{\Lambda^g}\widehat{\Phi^g}\widehat{\Lambda^{g'}} + \widehat{\Theta^g}$. This solution, in turn, produces $\widehat{\Sigma^g}$, which contains the estimates of population covariances among x variables under the assumption that a hypothesized factor structure holds in the population.

2.3. Estimation

Assuming that observed responses follow a multivariate normal distribution in the population, the ML estimates of the parameters in Equations 9 and 10 are obtained by minimizing the discrepancy function

$$F_{ML}\left(S, \hat{\Sigma}\right) = -2 \ln L = \sum_{g=1}^{G} \left(\frac{N^g}{N}\right) f_{ML}\left(S, \hat{\Sigma}\right)^g, \quad (11)$$

where N^g is the number of observations in group g and N is the number of total observations across groups. The function $f_{ML}\left(S, \hat{\Sigma}\right)^g$ is further written as

$$f_{ML}\left(S, \hat{\Sigma}\right)^g = \ln\left|\widehat{\Sigma^g}\right| + tr\left(S^g\widehat{\Sigma^g}^{'}\right) - \ln|S^g| - p, \quad (12)$$

where

$$S^g = \left(\frac{1}{N^g}\right)\sum_{g=1}^{N^g}(\bar{x}^g - \tau^g - \Lambda^g\kappa^g)(\bar{x}^g - \tau^g - \Lambda^g\kappa^g)^{'}. \quad (13)$$

As observed in Equation 11, the ML discrepancy function is inversely related to the likelihood function. Given the data, therefore, the parameter estimates are those that minimize the discrepancy between *S* and $\hat{\Sigma}$ (or maximize the likelihood of the data) under a hypothesized model.

2.4. Model Fit

A critical issue in CFA is how to determine whether a particular model adequately fits the data. As Marsh (1994) noted, first of all, researchers need to ensure that the (iterative) estimation procedure converges to a proper solution (e.g., positively defined matrices, no out-of-range values, reasonable standard errors, etc.) and that the parameter estimates are reasonable in relation to a prior model as well as to common sense. For simplicity, it is presumed here that these prerequisites have been satisfied for a particular MACS model.

The overall fit of a model is based on the discrepancy between the observed covariance matrix and reconstructed population covariance matrix. It is also based on the discrepancy between the observed mean vector and reconstructed population mean vector (Sörbom, 1974). The null hypothesis (H_0) for testing a particular model is that the hypothesized factor structure in the model holds exactly in the population. It can be written as

$$H_0: \Sigma = \Lambda\Phi\Lambda' + \Theta,$$

$$\mu = \tau + \Lambda\kappa.$$

The alternative hypothesis (H_a) is that Σ has no particular structure. It should be noted that the role of null hypothesis is reversed from its usual role in research.

Thus, failure to reject H_0 implies that the hypothesized model is plausible in the population.

The conventional measure of overall model fit is the chi-square statistic (Jöreskog, 1971). Under the null hypothesis, the ML discrepancy function (Equation 11) value yields a test statistic

$$(N-1)F_{ML}\left(S,\hat{\Sigma}\right), \quad (14)$$

which follows a chi-square distribution as *N* becomes large, with degrees of freedom

$$df = \frac{1}{2}p(p+3) - \left\{p - pm - \frac{1}{2}m(m-1)\right\}. \quad (15)$$

If the chi-square value is significant, we reject H_0. Otherwise, we cannot reject H_0; we have failed to show that a hypothesized model does not hold exactly in the population, thereby concluding that this model is tenable.

Although the chi-square statistic is the most commonly used measure of overall fit in the literature, many researchers have been concerned about its appropriateness (e.g., Bentler, 1990; Bentler & Bonett, 1980; Browne & Cudeck, 1993; Cudeck & Browne, 1983; Jöreskog & Sörbom, 1989). First, the conclusions based on the chi-square test can vary depending on *N* (see Equation 14); when *N* is sufficiently large, any parsimonious model will be rejected. Second, this test is extremely sensitive to (small to moderate) deviation from normality of the data. It is presumed that responses are multivariate normally distributed in each group. However, a distributional violation can occur when dichotomous or polytomous responses are analyzed. In a simulation study, West, Finch, and Curran (1995) showed that the chi-square test tends to reject H_0 for polytomous responses even

when their discrepancy function is small. Finally, the null hypothesis of perfect fit is a priori false when applied to real data (Marsh, Balla, & McDonald, 1988). Thus, the chi-square test essentially tests whether sample size is large enough for the test to tell us what we already know.

Alternatively, a large number of practical goodness-of-fit measures have been proposed in the literature. Those most commonly used are the Comparative Fit Index (CFI; Bentler, 1990), Non-Normed Fit Index (NNFI; Bentler & Bonett, 1980), and Root Mean Squared Error of Approximation (RMSEA; Steiger & Lind, 1980). Because most of the measures do not have a known sampling distribution, researchers recommend certain criterion values indicative of satisfactory model fit. Thus, it has been a common practice to report multiple goodness-of-fit measures when researchers evaluate a proposed model (Hu & Bentler, 1999).

2.5. Measurement Invariance

In CFA, measurement equivalence is referred to as measurement invariance. In his landmark work, Meredith (1993) used Lawley's (1943-44) selection theorem as a theoretical framework for measurement invariance. That is, if a particular factor structure holds in a population, the same structure should hold in any samples of the population no matter how they are chosen. Nevertheless, selection may introduce some dependency among unique factor scores and/or between unique factor scores and latent trait scores. Thus, a scale is said to have measurement invariance when conditional distributions of item responses are identical across groups, given a value on the trait (Meredith & Teresi, 2006).

Vandenberg and Lance (2000) extensively reviewed different levels of measurement invariance proposed in the literature and recommended a number of invariance tests that could be performed in empirical research. Moreover, Vandenberg (2002) illustrated how different invariance levels are required to answer different research questions (see also Steenkamp & Baumgartner, 1998).

2.5.1. Full Measurement Invariance

Configural Invariance

Configural invariance is based on Thurstone's (1947) principle of simple structure (Horn & McArdle, 1992; Horn, McArdle, & Mason, 1983). That is, items of a scale should exhibit the same pattern of salient (non-zero) and non-salient (zero or near zero) loadings across groups. Although, in principle, it is not necessary to constrain the non-salient loadings to zero, this is commonly done in CFA (Steenkamp & Baumgartner, 1998). As such, configural invariance requires only that the same number of latent trait variables and the same pattern of zero and salient loadings are specified in each group. It should be noted that no equality constraints are imposed on the parameters. Configural invariance is established by testing the null hypothesis that covariance and mean structures are equal across groups,

$$H_0: \Sigma^g = \Lambda^g \Phi^g \Lambda^{g'} + \Theta^g,$$

$$\mu^g = \tau^g + \Lambda^g \kappa^g \text{ for all } g.$$

Metric Invariance

Metric invariance introduces the concept of equal unit of measurement. If an item satisfies metric invariance, observed item responses can be meaningfully

compared across groups. Furthermore, comparisons of the latent trait variances and covariances become plausible.

Because the loadings carry information about how changes in the trait scores relate to changes in the observed scores, metric invariance is established by testing the null hypothesis that loadings are equal across groups,

$$H_0: \Lambda^1 = \Lambda^2 = \cdots = \Lambda^G.$$

Metric invariance does not necessarily indicate that the origins of the scale are equivalent across groups. Consequently, mean comparisons are not tenable yet, thereby leading Meredith (1993) to categorize this level of invariance as weak factorial invariance.

Scalar Invariance

Scalar invariance addresses the question of whether the latent trait mean differences are consistent with the observed mean differences (Steenkamp & Baumgartner, 1998). Even if an item satisfies metric invariance, scores on that item can still be systematically biased upward or downward (i.e., additive bias; Meredith, 1995). Given scalar invariance, researchers can ascertain whether the origins of the scale, as well as the unit of measurement, are identical across groups. As a consequence, either observed mean or trait mean comparisons become meaningful, thereby leading Meredith (1993) to term this level of invariance as strong factorial invariance.

If metric invariance has been satisfied, scalar invariance is established by testing the null hypothesis that intercepts are equal across groups,

$$H_0: \tau^1 = \tau^2 = \cdots = \tau^G.$$

Invariance of Unique Factor Variances

A final invariance level that may be imposed on the measurement model is that the unique factor variances are invariant across groups. If metric and scalar invariance have been satisfied, the invariance of unique factor variances is established by testing the null hypothesis specifying

$$H_0: \Theta^1 = \Theta^2 = \cdots = \Theta^G.$$

This level of invariance implies that "all group differences on the measured variables are captured by, and attributable to, group differences on the common factors" (Widaman & Reise, 1997). Meredith (1993) classified this level of invariance as "strict" factorial invariance.

In reality, the invariance of unique factor variances is extremely difficult to achieve. Widaman and Reise (1997) argued that the unique factor variances are not necessarily identical in practical applications and only metric and scalar invariance is essential for answering most research questions. However, when comparisons of the observed associations (e.g., correlation) are the questions of interest, reliability of the measure should be about the same in order for measurement artifacts not to bias the conclusions (Steenkamp & Baumgartner, 1998). "Reliability equality" is established if items of a test satisfy metric invariance, and only if the invariance of unique factor variances is defensible (Byrne, 1998).

Invariance of Factor Variances/Covariances and Factor Means

The invariance levels often imposed on the structural model are that the factor variances and/or factor means are invariant across groups. These invariance levels are evaluated if the previous invariance levels imposed on the measurement model have been satisfied.

The invariance of factor covariances is established by testing the null hypothesis that latent trait covariances are equal across groups,

$$H_0: \Phi_{jk}^1 = \Phi_{jk}^2 = \cdots = \Phi_{jk}^G \ (j = 2, \ldots, m; k = 1, \ldots, [j - 1]).$$

The invariance of factor variances is supported by testing the null hypothesis that trait variances are equal across groups,

$$H_0: \Phi_{jj}^1 = \Phi_{jj}^2 = \cdots = \Phi_{jj}^G \ (j = 1, \ldots, m).$$

If the invariance of factor variances and covariances is satisfied, the trait correlations are considered to be invariant across groups.

The invariance of factor means is established by testing the null hypothesis that latent trait means are equal across groups,

$$H_0: \kappa^1 = \kappa^2 = \cdots = \kappa^G.$$

The nonequivalence of the trait means is generally referred to as item impact in the DIF literature (Raju et al., 2002).

2.5.2. Partial Measurement Invariance

The aforementioned invariance tests are omnibus tests in the sense that they address the question of whether imposed equality constraints are fully satisfied. For example, metric invariance requires that all the loadings to be invariant across groups.

Muthén and Christoffersson (1981), however, implied that it is possible to test metric invariance when only some of the loadings were invariant. They termed this "partial" measurement invariance.

Byrne et al. (1989) provided a didactic article on how to test the level of partial measurement invariance. The basic idea is that full invariance is not necessary in order for further invariance tests and substantive analyses to be conducted (see also Meredith, 1993). In particular, they proposed that mean comparisons would be meaningful if metric and scalar invariance have been satisfied for at least two items per latent trait. A test for trait mean differences is supposedly more beneficial than one for observed mean differences because measurement error has been partialed out from the trait means. Furthermore, the trait mean differences will be estimated more accurately with imposed partial invariance constraints because the trait mean estimates are adjusted for the fact that only partial, not full, invariance characterizes the data (Cheung & Rensvold, 2000). However, one limitation is that the trait being compared may have different meanings for different groups under partial measurement invariance.

2.5.3. Testing Procedure

The procedure for invariance tests starts with an omnibus test that evaluates the equality of observed covariance matrices and mean vectors, both separately and jointly (Steenkamp & Baumgartner, 1998). In the unlikely cases that observed covariances and means are actually invariant across groups, analysis for separate groups is no longer necessary (i.e., data can be pooled). However, the omnibus test

has undergone some criticism. For example, Muthén (cited in Raju et al., 2002) and Rock, Werts, and Flaugher (1978) showed that this test can signify equal covariance matrices and mean vectors even when more specific invariance tests find otherwise. Furthermore, Byrne (1998) argued that the omnibus test should not be regarded as a necessary prerequisite to more specific invariance tests .Thus, regardless of whether or not the omnibus test indicates a lack of invariance, subsequent tests are recommended in order to pinpoint possible sources of noninvariance (Meade & Lautenschlager, 2004).

Thus, the model of configural invariance serves as a baseline model in the invariance tests (Horn & McArdle, 1992). Given that a baseline model represents the best model in terms of both parsimony and meaningfulness, it is possible that the baseline model may not be completely identical across groups (Byrne et al., 1989). Even if this is the case, subsequent invariance tests still continue by implementing a condition of partial invariance. For only those latent trait variables that support configural invariance, metric invariance is tested. Those loadings that do not conform to metric invariance remain unconstrained in the subsequent tests. Next, scalar or partial scalar invariance is tested only if at least partial metric invariance has been established. Similarly, those intercepts that do not conform to scalar invariance remain unconstrained in the subsequent tests.

The order of the invariance tests for unique factor variances, factor covariances, and factor variances is somewhat arbitrary (Bollen, 1989; Jöreskog, 1971). Ultimately, the order may not be critical in the sense that a particular level of

invariance is not required in order for subsequent invariance tests to be conducted. Indeed, often the invariance of factor covariances and factor variances is tested simultaneously (i.e., $\Phi^1 = \Phi^2 = \cdots = \Phi^G$). When the invariance of unique factor variances is examined, the test proceeds only if at least partial metric and scalar invariance has been established. At this point, those intercepts that do not conform to scalar invariance are unconstrained across groups.

2.6. Significance Test

The statistical framework for invariance tests was originally developed by Jöreskog (1971). The invariance tests require a series of hierarchically nested models to be estimated; a hypothesized model in which parameters of interest are constrained to be equal across groups is compared with a competing, less restrictive model in which the same parameters are freely estimated in each group. A particular level of invariance is satisfied if the model fit is adequate, and if its difference from the competing model is minimal (Widaman & Reise, 1997). Likewise, the same criteria hold in testing all subsequent invariance levels.

The standard way to compare the fit of two nested models is the LR test (Jöreskog, 1971). Consider two hypotheses that specify increasingly restrictive models, H_a and H_0. Let F_{MLa} and F_{ML0} be the minimum values of the ML discrepancy function under H_a and H_0, respectively. Under H_0, the test statistic

$$D = n(F_{ML0} - F_{MLa}), \quad (16)$$

where $n = N - 1$, follows asymptotically a chi-square distribution, with degrees of freedom equal to the difference in the degrees of freedom between the two models. If

the chi-square difference value is significant, we reject H_0; we conclude that the constraints specified in the more restrictive model do not hold. Otherwise, we conclude that all the equality constraints are tenable.

Steiger et al. (1985) noted that the LR test is quite flexible; it can test multiple constraints simultaneously, and when a series of the LR tests is conducted on a sequence of nested models, they are asymptotically independent. In a simulation study, Meade and Lautenschlager (2004) examined the power of the LR test in multiple-group cases. They showed that this test is fairly effective under optimal conditions. Not only was the omnibus test for equal observed covariance matrices successful in general, but also the lack of full metric invariance was accurately detected in most conditions.

On the other hand, several researchers (e.g., Cheung & Rensvold, 2002; Little, 1997; Marsh, Hey, & Roche, 1997; West, Finch, & Curran, 1995) have argued that the LR chi-square value is as sensitive to sample size and nonnormality of data. Theoretically, this statistic holds whether or not a baseline model is misspecified (Steiger et al., 1985). However, Yuan and Bentler (2004) found that it is an unreliable measure of relative model fit when a baseline model, in fact, has been misspecified. Kaplan (1989) also found that the power of the LR test, under partial metric invariance, is dependent on the size of the misspecification as well as on the correlation between the misspecified parameter and the remaining parameters in the model.

Alternatively, Cheung and Rensvold (2002) examined properties of 20 other goodness-of-fit measures proposed in the literature. Their simulation results indicated that each of ΔCFI, ΔGamma Hat (Steiger, 1989), and ΔNon-Centrality Index (NCI; McDonald, 1989) controlled its Type I error at the nominal alpha level (e.g., .05) when used to test (full) invariance. For ΔCFI, they suggested that the null hypothesis should not be rejected with a value smaller than or equal to –0.01. For ΔGamma Hat and ΔNCI, the critical values were suggested to be –0.001 and –0.02, respectively. In a recent conservative simulation study (i.e., .90 power, .01 Type I error), Meade, Johnson, and Braddy (2008) recommended that a change in CFI of more than 0.002 is the optimal criterion for rejecting the null hypothesis of invariance. Little (in press) suggested that, for most applications, Cheung and Rensvold's (2002) ΔCFI value of –0.01 is quite reasonable, but that Meade et al.'s (2008) ΔCFI value 0f –0.002 can be used when the question warrants such a restrictive criterion (e.g., high-stake testing environments). For example, Byrne and Stewart (2006) evaluated scalar invariance for the Beck Depression Inventory II (Beck, Steer, & Brown, 1996) among Hong Kong and American adolescents, using the Cheung and Rensvold's criterion.

2.7. Scaling

In any CFA model, indeterminacy exists between the scale of the item parameters and the scale of the latent trait variables. That is, if the scale for the item parameters should be identified to obtain unique estimates, the scale for the trait variables should be specified, or vise versa. The scaling is typically achieved by imposing a set of constraints on the parameters (Jöreskog & Sörbom, 1989).

Recently, Little, Slegers, and Card (2006) outlined two primary scaling methods for both mean (Equation 8) and covariance (Equation 9) structures. They also proposed another possible, statistically equivalent scaling method, called the *effects-coded method*. All three of the scaling methods yield identical statistical fit values because they are simple reparameterizations of one another (Little et al., 2006).

When three or more items are used to measure a latent trait variable, each method provides the necessary condition for identifying the scale of the trait variable. Consequently, they can be used for either single-group cases or multiple-group cases. However, having fewer than three items is problematic as a general rule. Thus, our discussion will focus on situations when a researcher has three or more items per trait variable.

2.7.1. Marker-Variable Method (Bollen, 1989; Jöreskog & Sörbom, 1993)

The *marker-variable method* involves fixing the loading of an item (i.e., marker or anchor) to 1 and the intercept of this item to 0 for each latent trait variable. Then, trait means and variances are freely estimated in all groups. Consequently, this method sets the scales of the trait variables to be equivalent to those of the chosen anchor item.

However, this method has an undesirable property. The estimated trait means and variances can vary depending on which item is chosen as an anchor. Nevertheless, the choice is somewhat arbitrary because there is no absolute rule yet in the literature (Little et al., 2006).

2.7.2. Fixed-Factor Method (McArdle & McDonald, 1984)

The *fixed-factor method* involves fixing the mean and variance of each latent trait variable in the first group. With loadings and intercepts equated across groups, the trait means and variances are freely estimated in the subsequent groups. Consequently, the trait variables are scaled relative to the fixed trait mean and variance in the first group. In general, the trait mean is fixed to 0 (i.e., $\kappa_m^1 = 0$) and the trait variance is fixed to 1 (i.e., $\Phi_{mm}^1 = 1$). This choice results in placing the trait variables of the first group in correlation metric (Little et al., 2006).

The trait means and variances in the subsequent groups are also fixed in the model for configural invariance. Unlike other methods, this method needs to subsequently free the trait variances and/or means in latter groups when further invariance levels are tested.

2.7.3. Effects-Coded Method (Little et al., 2006)

The effects-coded method involves, for each latent trait variable, constraining the set of intercepts to sum to 0. It also constrains the set of loadings to average to 1, which is the same as requiring them to sum to the number of unique items. These constraints can be written as

$$\sum_{i=1}^{P} \tau_{im}^{g} = 0 \text{ and } \sum_{i=1}^{P} \lambda_{im}^{g} = P, \quad (17)$$

where $i = 1$ to P refers to summation across the set of P unique items for a given trait variable. The intercept parameters are estimated as an optimal balance around 0, but no individual intercept needs to be fixed. Similarly, the loading parameters are

estimated as an optimal balance around 1, but no particular loading is necessarily constrained.

The trait means and variances reflect the observed metric of the items, optimally weighted by the degree to which each item represents the underlying trait variable. Consequently, a given trait variable will be on the same scale as the average of all the items. This method is desirable in the sense that the average of a set of items would be a more accurate estimate of a population value than any one item arbitrarily chosen from the set (Little et al., 2006).

2.8. Applicability of CFA

CFA involves many practical features with regard to its application. First, CFA programs (e.g., AMOS, EQS, LISREL, Mplus) provide a number of "useful" measures of model fit, but IRT programs (e.g., BILOG, MULTILOG, PARSCALE) yield only the LR test statistic as a standard. Second, CFA allows researchers to work with responses on multidimensional questionnaires (see Little, 1997), measured in multiple groups. In contrast, many of the IRT programs assess measurement equivalence between only two groups and they are confined to unidimensional questionnaires. Given the advances by Kim, Cohen, and Park (1995) and Oshima, Raju, and Flowers (1997), however, IRT analysis will become possible for cases involving multiple ability dimensions measured in more than two groups.

2.9. CFA Methodologies for DIF Detection

The relationship between CFA and two-parameter IRT models for dichotomous responses was clarified by Takane and de Leeuw (1987) and McDonald

(1999). That is, the loading (λ_i) and intercept (τ_i) parameters in CFA are essentially equal to the discrimination (a_i) and attractiveness (b_i) parameters in IRT, respectively. Since then, the concept of item invariance has been integrated into a more general, theoretical framework provided by IRT. Consequently, researchers are now able to test DIF within CFA, or more broadly within the SEM framework. The CFA- and SEM-based techniques for DIF detection employ the MACS model (Chan, 2000; Ferrando, 1996) and the multiple indicators multiple causes (MIMIC) model (Muthén, 1988), respectively. The MIMIC technique is briefly illustrated, followed by a detailed discussion of the MACS technique. More details for the MIMIC technique can be found in Muthén (1988).

3. MIMIC Technique

The MIMIC model (Jöreskog & Goldberger, 1975) regresses the latent trait variables on exogenous observed variables (covariates). Muthén (1988) further extended this model such that the item responses are also regressed on the covariates. Thus, the observed response x_i to an item i (i = 1, …, p) is represented as a linear function of an intercept τ_i, trait variables ξ_j (j = 1, …, m), observed covariates z_c (c = 1, …, r), and a unique factor score δ_i. It can be written as

$$x_i = \tau_i + \lambda_{ij}\xi_i + \beta_{ic}z_c + \delta_i, \quad (18)$$

where β_{ic} is a $p \times r$ matrix of regression slopes that represent the effects of the covariates on the item responses. Under the usual assumptions, taking the expectation of Equation 18 yields the relation between the observed item means and the latent trait means:

$$\mu = \tau + \Lambda\kappa + \mathrm{B}z, \quad (19)$$

where μ is a $p \times 1$ vector of item means, κ is an $m \times 1$ vector of trait means, and B is a $p \times r$ matrix of regression slopes of the item responses on the covariates.

The regression slopes in B are called the *direct effects* because they influence the responses, unmediated by the latent traits. The direct effects indicate whether the item responses differ across groups after controlling for any trait mean differences, which is the definition of DIF (Fleishman, 2005; Fleishman & Lawrence, 2003; Fleishman, Spector, & Altman, 2002; Millsap & Everson, 1993). Accordingly, DIF is evident when the direct effects are statistically significant (Grayson, Mackinnon, Jorm, Creasey, & Broe, 2000; Jones, 2006). Because the loadings are assumed to be equal across groups, the MIMIC technique is limited to tests for uniform DIF.

4. MACS Technique

4.1. Specification

The standard MACS model can be extended for DIF detection by addressing a number of additional assumptions. Assuming that only "one" latent trait variable accounts for continuous responses on a scale (i.e., congeneric item responses; Jöreskog, 1971), observed responses *x* to an item *i* (*i* = 1, …, *p*) are explained by means of linear regression on the trait variable ξ in this particular MACS model. More specifically,

$$x_i = \tau_i + \lambda_i\xi + \delta_i. \quad (20)$$

As noted in Equation 6, the intercept τ_i represents the expected response to an item *i* for examinees with trait scores of zero. The factor loading λ_i refers to the expected

change in the item response per unit change in the trait variable. Finally, δ_i is the unique factor score, which is assumed to be normally distributed. Under the further assumption that the covariances among the unique factor scores are zero in the population, the mean of x_i is equal to τ_i when the trait score is zero and covariances between x_i and ξ are equal to λ_i (Jöreskog, 1971). Thus, taking the expectation of Equation 20 yields the covariance matrix of *x* variables in the population

$$\Sigma = \Lambda\Phi\Lambda' + \Theta, \quad (21)$$

where Λ is a $p \times 1$ vector of factor loadings and Θ is a $p \times p$ diagonal matrix of unique factor score variances. The mean vector of *x* variables in the population is given by

$$\mu = \tau + \Lambda\kappa, \quad (22)$$

where τ is a $p \times 1$ vector of intercepts and κ is a scalar trait mean.

The assumptions that (a) a single latent trait underlies the correlations among the observed responses and that (b) off-diagonal elements in Θ_g are zero are the analogs of, respectively, the unidimensionality and local independence[2] assumptions in IRT. In the context of IRT, τ_i corresponds to the attractiveness parameter (i.e., the observed mean for examinees with zero trait score) and λ_i the discrimination parameter (i.e., the ability of an item to differentiate examinees with different trait scores) (Grayson & Marsh, 1994; Mellenbergh, 1994). Under the assumption that the same factor structure underlies each group *g* (*g* = 1, …, *G*), Equations 21 and 22 are extended to

[2] Note that the local independence assumption can be violated and estimated by specifying the correlated true population residuals.

$$\Sigma^g = \Lambda^g \Phi^g \Lambda^{g'} + \Theta^g \text{ and } \mu^g = \tau^g + \Lambda^g \kappa^g, \quad (23)$$

respectively.

4.2. Relation to IRT

The MACS model for congeneric responses and the IRT model for dichotomous responses assume different conditional distributions of the responses (i.e., normal with homogeneous variance and binomial, respectively). But, Mellenbergh (1994) noted that the MACS model has the same structure as the two-parameter dichotomous IRT model (i.e., Birnbaum, 1968). Thus, they hold two invariance properties in common; the invariance of item parameters over population samples and the invariance of an examinee's latent score over measurements (see Hambleton & Van der Linden, 1982).

Although the latent trait variable is believed to have a mean of zero and a variance of unity in the population, this property does not necessarily hold in the samples from the population. That is, the trait variable will not usually have the zero mean and unity variance in samples because the covariances among unique factor scores are not expected to be zero (MacCallum & Tucker, 1991; Meredith, 1993). Thus, it can be assumed (and tested) that item parameters are invariant but trait means and variances can vary from those in the population (Ferrando, 1996).

The intercepts in MACS analysis correspond to the attractiveness/difficulty parameters in IRT; the higher the intercept, the more attractive/difficult the item is (i.e., a higher mean response is obtained). Factor loadings correspond to the discrimination parameters; the higher the loading, the more discriminating the item is

(i.e., examinees of different latent scores are better differentiated; see Ferrando, 1996; Grayson & Marsh, 1994; Mellenbergh, 1994).

The intercept parameter represents differential response level associated with an item, whereas the loading parameter represents how concretely the item reflects the trait variable being measured (Ferrando, 1996). Thus, the intercept parameter corresponds to the attractiveness parameter in IRT; the higher the intercept, the more attractive the item is in the sense that a higher mean response is obtained. Similarly, the loading parameter corresponds to the discrimination parameter in IRT; the higher the loading, the better the item differentiates examinees with different trait scores (see Ferrando, 1996; Grayson & Marsh, 1994; Mellenbergh, 1994).

As noted previously, uniform DIF exists when only the attractiveness (intercept) parameter differs across groups. Non-uniform DIF is present when the discrimination (loading) parameter differs across groups, regardless of whether or not the attractiveness parameter is invariant. Thus, lack of invariance in τ^g implies uniform DIF, while lack of invariance in Λ^g implies non-uniform DIF, regardless of the invariance in τ^g (Chan, 2000).

The unique factor score provides information about precision in measurement. Indeed, the *item information function* (IIF) in IRT is equal to the ratio of the squared loading and unique factor score variance (Mellenbergh, 1994). However, the unique factor score is not a parameter of substantive interest in the DIF literature. As such, invariance of unique factor variances is not usually of concern in the MACS technique for DIF analysis.

The illustrated MACS model has been specified only for continuous responses. But, it is also applicable to the case of items with dichotomous or ordered response options. For these items, it is assumed that the response options correspond to segments of a latent continuous response variable (Mellenbergh, 1994) denoted by x_i^*. Thus, the MACS model for the latent response variable can be written as

$$x_i^* = \tau_i + \lambda_i \xi + \delta_i, \quad (24)$$

Then, the threshold parameter γ is introduced to accommodate the discrete nature of the observed dichotomous and polytomous (*k*-category) responses:

$$x_i = \begin{cases} 1 \text{ if } x_i^* \geq \gamma_i \\ 0 \text{ if } x_i^* < \gamma_i \end{cases} \text{ and } x_i = k \text{ if } \gamma_{ik} < x_i^* \leq \gamma_{ik+1}, \quad (25)$$

respectively. Because the latent continuous responses are assumed to be multivariate normally distributed in the population, a polychoric correlation matrix of the observed responses is computed (Bollen, 1989) and then the MACS model is fitted to the matrix (e.g., Jones, 2004; Stark et al., 2006).

As mentioned previously, a direct correspondence exists between IRT item parameters and CFA item parameters in the case of binary items. Specifically, implementing the two-parameter IRT model and the parameter standardization of the CFA model using, say, trait mean of zero and variance of unity and unity variance of the latent response variables,

$$a_i = \lambda_i / \sqrt{1 - \lambda_i^2} \text{ and } b_i = \gamma_i / \sqrt{1 - \lambda_i^2}. \quad (26)$$

The conditional probability of the observed response to an item *i* then can be obtained as

$$P_i(x_i = 1|\xi_j) = 1 - \Phi\left(\gamma_i - \lambda_i\xi_j \Big/ \sqrt{1 - \lambda_i^2}\right), \quad (27)$$

where $P_i(\xi_j)$ is the probability that an examinee j with a given value on ξ will correctly answer an item *i* and Φ is the cumulative standard normal distribution function. A complete explication of these relationships can be found in Kamata and Bauer (2008), Lord and Novick (1968), McDonald (1999), Muthén and Christoffersson (1981), and Takane and de Leeuw (1987).

4.3. Strategy

In the CFA literature, the MACS technique has been applied for testing both uniform and non-uniform DIF, using either a "constrained-baseline" strategy or a "free-baseline" strategy. Regardless of which strategy is used, tests for uniform DIF are typically conducted only for those items that have been found to have no non-uniform DIF.[3] This two-step procedure is consistent with the fact that metric (loading) invariance is cited as a prerequisite for scalar (intercept) invariance (Vandenberg, 2002; Vandenberg & Lance, 2000).

4.3.1. Constrained-Baseline Strategy

The constrained-baseline strategy tests for DIF one item at a time, assuming that other items are DIF-free anchors (e.g., Chan, 2000; Chen & Anthony, 2003; Finch, 2005; Gelin, 2005; Muthén & Asparouhov, 2002; Oishi, 2006; see Stark, et al., 2006). This strategy starts with a "fully-constrained" baseline model (Model A), in which all the loadings and all the intercepts are constrained to be equal across groups.

[3] Non-uniform DIF (loading invariance) and uniform DIF (intercept invariance) for a particular item can be tested simultaneously (see Stark et al., 2006).

After overall fit of this model is established, it is then statistically compared with each of *p* models (where *p* = number of items), in which one respective loading is freely estimated in each group. Next, a nested baseline model (Model B) is fitted, in which the loadings of the (previously identified) non-uniform DIF items are allowed to vary across groups. This model is then compared with each of *q* models (where *q* = number of items with invariant loadings), in which one respective intercept is freely estimated in each group.

The statistical significance of the parameter (i.e., loading, intercept) invariance is usually determined by conducting the modification index (MI) test or LR test. The MI indicates how much the LR chi-square value is "likely" to reduce if a particular fixed parameter is freely estimated. Thus, the MI values associated with the loadings and intercepts are obtained from the baseline Models A and B, respectively. The critical chi-square value for 1 degree of freedom is used as a MI criterion value for flagging DIF. Alternatively, DIF is indicated when the LR test statistic is statistically significant in a series of "actual" nested-model comparisons. A Bonferroni correction is recommended to set the critical (MI and chi-square) values at a reduced alpha level (i.e., $p = \alpha$ / number of invariance tests).

In a couple of simulation studies (González-Romá, et al., 2006; Hernández & González-Romá, 2003), the constrained-baseline strategy was found to perform fairly well; it maintained reasonable control for Type I error and satisfactory power in the medium to large DIF conditions.

4.3.2. Free-Baseline Strategy

The free-baseline strategy tests DIF in each item separately, assuming that other items are not free from DIF (e.g., Fleishman et al., 2002; Woods, 2009; Woods, Oltmanns, & Turkheimer, 2008; see Stark et al., 2006). This strategy starts with a "fully-free" baseline model (Model A; see Figure 1A), in which all parameters are freely estimated in each group except those needed for scaling. Once overall model fit is established, this model is then statistically compared with each of *p* models that constrain one respective loading to be equal across groups (Model B; see Figure 1B). Next, a nested baseline model (Model C) is fitted, in which the (previously identified) invariant loadings are constrained to be equal. This model is then compared with each of *q* models (Model D; see Figure 1C), in which one respective intercept is constrained to be equal across groups. Non-uniform DIF is indicated if the chi-square difference between Models A and B is statistically significant with 1 degree of freedom. Similarly, if the chi-square difference between Models C and D is significant, this item is considered to exhibit uniform DIF. A Bonferroni correction for multiple nested-model comparisons is also recommended for this strategy.

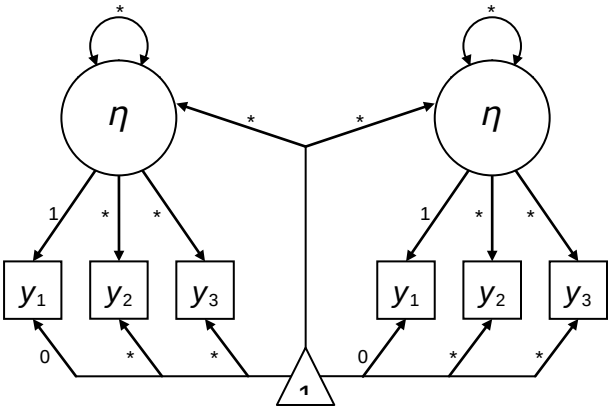
Figure 1

Free-Baseline Strategy for the MACS Technique for DIF Detection

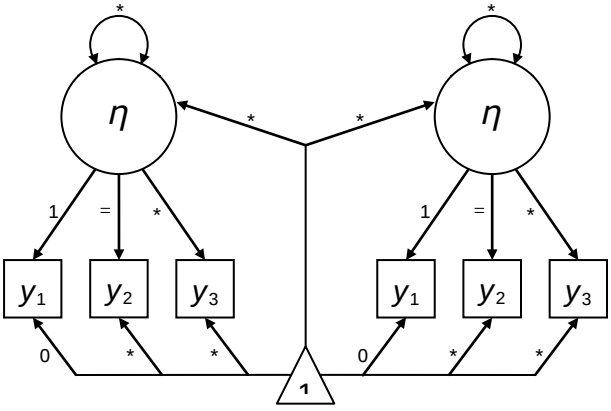
Figure 1A to 1C depict three nested MACS models. This example illustrates a simple case, in which (a) the scale includes three items, (b) only the second item exhibits non-uniform DIF, and (c) the marker-variable scaling method is used. For simplicity,

the unique factor variances are omitted. The free parameters are marked by "*" and the parameters equated across groups are marked by "=."

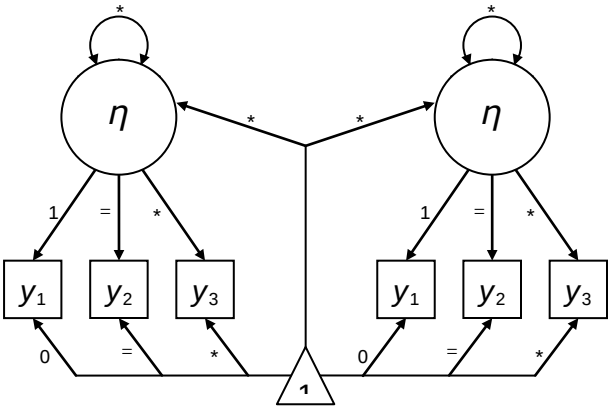
A. Fully-Free Baseline Model



B. Model of a Single Restrictive Loading



C. Model of a Set of Restrictive Loading and Intercept



In a simulation study, Stark et al. (2006) evaluated a variant of the free-baseline strategy, which tests uniform DIF and non-uniform DIF simultaneously. They found that the free-baseline strategy performs reasonably well in all study conditions; power was high while Type I error was acceptable at the nominal alpha level. When a Bonferroni correction was applied to the large sample, large DIF conditions, Type I error was almost eliminated while power remained high.

4.3.3. Recommended Strategy

Although previous simulation studies have been supportive, the constrained-baseline strategy entails apparent problems. First, the use of MI involves the danger of capitalizing on chance (Steenkamp & Baumgartner, 1998). That is, idiosyncrasies of a particular data set may necessitate revisions of a hypothesized model, which cannot be replicated with different data. Such data-driven MI may invalidate the probability values (i.e., Type I error rate and power) associated with subsequent LR tests (Gregorich, 2006). Indeed, MacCallum (1986) found that MI is particularly unsuccessful in uncovering a misspecified parameter (see also Kaplan, 1989; Luijben, Boomsma, & Molenaar, 1987).

Second, and more importantly, the constrained baseline model is not theoretically suitable for nested-model comparison. In order for the LR test statistics to follow a central chi-square distribution under the null hypothesis, the baseline model should fit the data adequately (Maydeu-Olivares & Cai, 2006). As mentioned previously, the constrained-baseline model assumes equal loadings and/or equal intercepts across groups. In the likely case that a scale includes one or more DIF

items, this model may not fit adequately (e.g., Marsh, Balla, & McDonald, 1988). Consequently, the MI test will be inaccurate or the LR test will be untenable (Cheung & Rensvold, 1999). Evidently, Stark et al. (2006) found that the constrained-baseline strategy performs well only when there is no DIF at all in a scale.

In contrast, the free-baseline strategy is theoretically suitable for DIF analysis in the sense that (a) DIF is tested for one item at a time, (b) the parameter estimates are not subject to relative size or significance of other parameters in the model, and (c) the baseline model reasonably provides a proper fit value, against which DIF is examined in the subsequent nested models. Thus, the free-baseline strategy is recommended, rather than the constrained-baseline strategy, when testing for DIF. Accordingly, the former strategy is followed in the present simulation study, which examines the performance of the MACS technique under various conditions.

5. Misspecification Problems

5.1. Invariance Assumption in the Conventional Scaling

The MACS analysis for DIF detection, which is a simple variation of the idea of partial measurement invariance (Byrne et al., 2002), involves some methodological issues to be resolved in practice. In general, the scaling does not change the conclusions about overall model fit or the tests for whether or not a given level of invariance holds (González & Griffin, 2001; Little et al., 2006). However, when a researcher locates DIF in a scale after metric or scalar invariance has been rejected, a potential problem arises. That is, different scaling methods can lead to different

outcomes of DIF tests because these post-hoc tests rely on an examination of individual parameters.

As noted previously, a set of constraints are imposed on the parameters for scaling. When there is more than one group, the scaling is conventionally achieved by constraining an item's (marker or anchor) parameters to be equal across groups (Vandenberg & Lance, 2000). When used for the (free-baseline) MACS analysis, this marker-variable scaling method fixes an anchor item's loading to 1 and intercept to 0 for all groups. This practice is essentially the same as assuming that the chosen anchor is truly invariant (Chan, 2000; see González-Romá et al., 2006). If the "biasedness" of the anchor cannot be guaranteed or a researcher chooses an anchor arbitrarily, other parameter estimates are placed on different scales across different groups (Bollen, 1989). This problem may account for more often reported problems of inflated Type I error (Cheung & Rensvold, 1999; see Finch, 2005; Meade & Lautenschlager, 2004; Stark et al., 2006; Wang, 2004; Wang & Yeh, 2003). For example, Stark et al. (2006) found that a biased anchor set severely inflates Type I error of the (constrained-baseline) MACS analysis for DIF detection. The inflation was greater as true differences in the anchor set parameters increased between groups. Accordingly, González-Romá et al. (2006) called for future research to examine the effects of using a DIF item as a single anchor.[4]

[4] More specifically, a single anchor is a sufficient condition for scaling if the conventional marker-variable scaling method is used with the free-baseline strategy.

5.2. Choice of an Unbiased Anchor

A designated, invariant anchor set is desirable for any DIF analysis. However, the designated anchor set is usually selected based on preliminary analyses of the same data that will be used for the main analyses, rather than based on extensive prior research (Woods, 2008; see Thissen et al., 1993). To rule out the possibility of bias in an anchor set, a variety of empirical solutions have been proposed in the literature. For example, Cheung and Rensvold (1999) and Wang (2004) suggested using all items once as an anchor. González-Romá et al. (2005) repeated the (free-baseline) MACS analysis, while randomly selecting an anchor. Nevertheless, such iterative solutions become quite labor intensive for many practitioners as the total number of items increases. In addition, Type I error will be severely inflated if no item appears to be invariant, even when the alpha level has been adjusted for the increased number of nested-model comparisons.

More recently, Stark et al. (2006) proposed a two-step process. That is, while running the constrained-baseline MACS analysis, the first step involves selecting an item that has the highest loading and is "(presumably) unbiased" (p. 1304). The second step involves conducting the free-baseline analysis, while using the selected item as an anchor in the subsequent DIF tests. However, this two-step process involves a couple of uncertainties. First, the baseline model used in this step is not theoretically reasonable (see *4.3.3. Recommended Strategy* in this dissertation) and therefore failure at the first step will jeopardize the validity of the second step for DIF analysis. Second, although an anchor should be highly related to the latent trait

variable because it defines the scale of the trait variable, there is no known relationship between the magnitude of the loading and the amount of DIF (Woods, 2009).

5.3. Potential Resolutions for Misspecification Problems

When the (free-baseline) MACS analysis is conducted with more than one group, scaling is achieved conventionally by constraining the item parameters of a chosen anchor to equality across groups. In contrast, alternative scaling methods do not impose a between-group equality constraint on a particular item. In other words, they do not require a researcher to have a designated anchor or anchor set.

Furthermore, the alternative scaling methods provide additional preferable features. For example, if the fixed-factor method is used, the latent trait variable is standardized so that item parameter estimates are readily convertible to those in the two-parameter IRT model (Kamata & Bauer, 2008; see Equations 26 and 27; see also McDonald, 1999; Takane & de Leeuw, 1987). Another advantage of using this method is that the association estimates among the trait variables have a correlation metric when more than one trait variable – each with a unique set of items – is modeled simultaneously. If the effects-coded method is used, item parameter estimates are optimally balanced so that the trait parameter estimates would be weighted, more accurate estimates of population values. The fact that, with either one of these alternative methods, all items of a scale can be tested for DIF may lead to more accurate conclusions about DIF. After a purification process (e.g., between-group equality of the trait parameters are evaluated, while imposing the constraints of

the supported partial invariance), group comparisons of the trait means and variances may become more meaningful.

6. Research Purpose

Despite the likely problems of model misspecification in the real world (Cheung & Rensvold, 1999), little empirical research has been conducted in the literature. Accordingly, the purpose of this dissertation is to present a new study which examines the effects of using a biased anchor, or more broadly the effects of using different scaling methods in the MACS DIF analysis.

It is important that a technique control its Type I error to be a valid statistical test of the hypothesis. If observed Type I error rates are found to be within reason, the power of the test needs to be examined. Accordingly, the Type I error and power of the MACS technique is explored by means of a Monte Carlo simulation.

7. Hypotheses

As noted previously, different but statistically equivalent scaling methods lead to the same conclusions about "omnibus" measurement invariance (González & Griffin, 2001; Little et al., 2006). But, they can yield different conclusions when the model tests DIF. The hypotheses of the present simulation study are as follows.

> *Hypothesis 1*: When the item parameters used for scaling are truly invariant across groups, the performance of the MACS technique will be equivalent, regardless of the scaling method.

More specifically, three different scaling methods will be comparable in terms of Type I error and power if the anchor is neither a uniform nor non-uniform DIF item.

Hypothesis 2: When the item parameters used for scaling are not truly invariant across groups, the performance of the MACS technique will depend on the choice of scaling method.

More specifically, if the anchor is either a uniform or a non-uniform DIF item, using the marker-variable method will inflate Type I error, consequently making the MACS technique unsuitable for testing DIF.

CHAPTER III: METHODOLOGY

This chapter discusses the methodology used for the present simulation study. A Monte Carlo simulation yielded Type I error and power for the mean and covariance structure (MACS) confirmatory factor analysis (CFA) technique to detect differential item functioning (DIF) under various conditions. The condition factors included those that have been frequently examined in the DIF literature.

Previous simulation studies have commonly manipulated type and amount of DIF, sample size or similarity of the sample sizes between focal and reference groups, latent trait distribution, type of item response, total number of items, and bias in the anchor set (see Hernández & González-Romá, 2003; González-Romá et al., 2006; Meade and Lautenschlager, 2004; Stark et al., 2006).[5] They found that the sample size, trait distribution, bias in an anchor set, and Bonferroni correction for multiple nested-model comparisons impact the Type I error in the MACS technique. Stark et al. (2006) found that moderate group differences in the latent trait mean (i.e., a 0.5 standard deviation difference) have little impact, whereas González-Romá et al. (2006) observed better control for Type I error when trait means and sample sizes were equal between groups. In terms of power, the previous simulation studies found that the MACS technique is positively related to the amount of DIF, sample size, and total number of items.

[5] Hernández & González-Romá (2003), González-Romá et al. (2006), and Meade and Lautenschlager (2004) used the constrained-baseline strategy, whereas Stark et al. (2006) used the free-baseline strategy as well.

Unfortunately, to my knowledge, there has been no simulation research considering the scaling method, which may affect the performance of the MACS technique. Furthermore, empirical evaluations of using different test statistics and criterion values are scant. In fact, French and Finch (2006) conducted a simulation study, where the conventional LR test and the ΔCFI (value of –0.01) test were compared separately for the case of testing omnibus metric invariance of a scale and the other case of testing non-uniform DIF in an item. When testing omnibus metric invariance with maximum likelihood (ML) estimation and normally distributed observed variables, the LR test maintained its Type I error at both .05 and .01 alpha levels in nearly all conditions. Under the same circumstance, the ΔCFI test provided comparable or less power than the LR test but inflated Type I error for small sample sizes. Power of detecting non-uniform DIF was reduced for both tests, but it was particularly low for the ΔCFI test. However, this study did not report Type I error for detecting non-uniform DIF. Further, this study did not consider locating uniform DIF in a scale.

Taken together, the condition factors manipulated in the present simulation study were as follows: type of item response, total number of items (scale size), similarity of sample size, similarity of latent trait mean, type of DIF in an anchor, type and amount of DIF in a target item, test statistic and criterion value, and scaling method.

1. Design

1.1. Type of Item response

The item responses were categorical with either two (i.e., dichotomous) or five options (i.e., polytomous). These categories are numbers frequently encountered in psychological tests and questionnaires, and five categories is the recommended minimum that adequately represents examinees' scores on ordinal items by means of the MACS model (Bollen & Barb, 1981; Dolan, 1994).

The item responses were conceptualized as an observed ordinal response *x*, wherein the underlying response *x** was latent and continuous (Mellenbergh, 1994). As the normally distributed latent response increases beyond certain threshold values, the observed response takes higher scores. Thus, an examinee who chooses one response category has more of a characteristic than another who chooses a lower category.

1.2. Scale size

The scale consisted of 6 or 12 items. The anchor, if required for scaling, was always Item 1, whereas Item 2 always served as a target item. When DIF was present, it appeared only on the anchor, only on the target item, or both. Consequently, the proportion of the DIF items in the scale ranged from 0 to .33.

1.3. Similarity of Sample Size

Three combinations of sample sizes were designed in this study; N_f = 100, 250, and 500 for a focal group and N_r = 900, 750, and 500 for a reference group.

Consequently, total sample sizes were always ($N_f + N_r$ =) 1,000, so as not to confound differences in the sample size with total sample size.

1.4. Similarity of Latent Trait Mean

The latent trait variable always followed a standard normal distribution ($\xi \sim N[0,1]$) in the reference group. The focal group had the same trait distribution, or the trait mean differed by 1 standard deviation so that this group had a smaller trait mean ($\xi \sim N[-1,1]$) than the reference group.

1.5. Type of Anchor

The anchor had no, non-uniform, or uniform DIF.[6]

1.6. Type of Target Item

Similarly, the target item had no, non-uniform, or uniform DIF. The DIF (anchor and target) items were less discriminative (non-uniform DIF) or less attractive (more difficult; uniform DIF) for examinees in the focal group.

1.7. Amount of DIF

For the target item (Item 2), the amount of non-uniform DIF was 0, 0.2, or 0.4 for no, small, and large DIF, respectively. Also, the amount of uniform DIF was 0, 0.3, or 0.8 for no, small, and large DIF, respectively. The amount of DIF in the anchor (Item 1), if present, was always large, regardless of whether it was non-uniform or uniform.

---

[6] It should be noted that both the fixed-factor and the effects-coded scaling methods do not require any anchor item. For these methods, the presence of DIF in the anchor should be interpreted as having DIF in one item (Item 1) when other, target item is being tested.

These values are comparable to those used in previous simulation studies; 0.6 (Finch, 2005), 0.1 and 0.2 (Kaplan & George, 1995), 0.25 (Meade & Lautenschlager, 2004), 0.75 (Navas-Ara & Gómez-Benito, 2002), 0.15 and 0.4 for non-uniform DIF and 0.25 and 0.5 for uniform DIF (Stark et al., 2006), 0.2, 0.5, and 0.8 (Wanichtanom, 2001), and 0.4 (Wang & Yeh, 2003).

1.8. Criterion Value

To determine the presence of DIF, the chi-square change and the CFI change were evaluated for each nested-model comparison. The *p* value for the LR chi-square test was uncorrected ($p = .05$) or Bonferroni-corrected ($p = .05$ / number of possible invariance tests). The critical ΔCFI values were –0.01 and –0.002 as proposed in the literature. When the observed test statistic was greater than the critical value, the target item was identified as having DIF.

1.9. Scaling Method

Three scaling methods were used; (a) the marker-variable method fixed the anchor's loading to 1 and intercept to 0 in each group, (b) the fixed-factor method fixed the latent trait mean and variance to 0 and 1 in each group, respectively, and (c) the effects-coded method constrained all the intercepts to average to 0 and all the loadings to average to 1 in each group.

All the condition factors were crossed with each other, resulting in a total of 5,184 conditions (two types of item response × two scale sizes × three combinations of sample sizes × two combinations of latent trait distributions × three types of anchor × three types of target item × two amounts of DIF in the target item × four criterion

values × three scaling methods). See Appendix A for a visual representation of the study design.

2. Data Generation

Both dichotomous and polytomous responses were generated in this study. First, population parameter values were specified such that the same factor structure underlay each group. The intercept difference of an item between two groups does not necessarily represents the mean difference of the item because the latter is influenced by loadings and latent trait means as well (Kamata & Bauer, 2008; Stark et al., 2006; see also Equation 8). In other words, raising the intercept of a particular item increases its attractiveness only when its loadings *and* the trait means are equal between two groups. To isolate the effects of varying item attractiveness and/or latent trait distribution on the performance of the MACS technique, therefore, a single common factor model was used for the response generation rather than the MACS model. The single common factor model can be written as follows:

$$x_i = \lambda_i \xi + \beta \delta_i, \quad (28)$$

where x_i represents the observed response *x* to an item *i*, λ_i represents the loading of x_i on a common factor ξ, and β represents the loading of x_i on a unique factor δ_i. The common factor loadings were equal between two groups, except those for the anchor (Item 1) and the target item (Item 2). The unique factor loadings were given by $\sqrt{1-\lambda_i^2}$, thereby yielding a variance of unity in the items. The population parameter values are shown in Tables 1 and 2.

The latent trait scores (ξ) were sampled from independent normal distributions (i.e., $\xi \sim N[0,1]$, $\xi \sim N[-1,1]$) and the unique factor scores (δ_i) were sampled from a standard normal distribution. They were substituted into Equation 28, along with the loadings previously specified, to create continuous item responses.

Once the continuous responses were obtained, they were divided into two or five ordered categories. For the dichotomous responses, an item threshold γ_i was chosen in accordance with 50% of the area under the normal curve. If a continuous response was greater than the item threshold $\gamma_i = 0$, then this response was scored as 1; Otherwise, it was scored as 0. For the polytomous conditions, four item thresholds of equal interval were chosen in accordance with approximately 3.6%, 23.8%, 45.1%, 23.8%, and 3.6% of the area under the normal curve. For each item, the ordinal responses (k) were assigned as such: $k = 1$ if $x_i \leq -1.8$; $k = 2$ if $-1.8 < x_i \leq -0.6$; $k = 3$ if $-0.6 < x_i \leq 0.6$; $k = 4$ if $0.6 < x_i \leq 1.8$; $k = 5$ if $x_i > 1.8$.

If present, DIF was created by varying an item's loading or threshold parameter between two groups. Specifically, the loading parameter for the focal group was reduced by a certain amount (i.e., 0.2, 0.4) to create non-uniform DIF. To create uniform DIF, the threshold parameter was raised by a certain amount (i.e., 0.3, 0.8). For polytomous responses, all threshold parameters were shifted by the same amount. This corresponds to varying all the attractive parameters obtained from graded response model (Samejima, 1979) between two groups.

Within each study condition, 500 replications were made in each group to avoid capitalizing on chance. This resulted in 648,000 data sets for analysis. The

same set of the trait scores was used but new values were sampled for the unique factor scores in each replication.

Table 1

Simulated Population Parameter Values for Binary Items

				Focal Group					
	Reference Group			DIF on λ			DIF on γ		
Item	λ	γ	β	λ	γ	β	λ	γ	β
1	.90	.00	.19	.50	.00	.75	.90	.80	.19
2	.80	.00	.36	.40 (.60)	.00	.84 (.64)	.80	.80 (.30)	.36
3	.75	.00	.44						
4	.70	.00	.51						
5	.65	.00	.58						
6	.55	.00	.70						
7	.95	.00	.10						
8	.85	.00	.28						
9	.45	.00	.80						
10	.35	.00	.88						
11	.30	.00	.91						
12	.25	.00	.94						

Note. Item 1 and 2 were used as the anchor and target items, respectively. The loading and threshold values in parentheses indicate that they used for the “small DIF” conditions; adjacent values were used for the “large DIF” conditions. The parameter values for the other items are not shown in the focal group because they were equal to those in the reference group.

Table 2

Simulated Population Parameter Values for Ordinal Items

	Reference Group						Focal Group DIF on λ					
Item	λ	γ_1	γ_2	γ_3	γ_4	β	λ	γ_1	γ_2	γ_3	γ_4	β
1	.90	-1.80	-.60	.60	1.80	.19	.50	-1.80	-.60	.60	1.80	.75
2	.80	-1.80	-.60	.60	1.80	.36	.40 (.60)	-1.80	-.60	.60	1.80	.84 (.64)
3	.75	-1.80	-.60	.60	1.80	.44						
4	.70	-1.80	-.60	.60	1.80	.51						
5	.65	-1.80	-.60	.60	1.80	.58						
6	.55	-1.80	-.60	.60	1.80	.70						
7	.95	-1.80	-.60	.60	1.80	.10						
8	.85	-1.80	-.60	.60	1.80	.28						
9	.45	-1.80	-.60	.60	1.80	.80						
10	.35	-1.80	-.60	.60	1.80	.88						
11	.30	-1.80	-.60	.60	1.80	.91						
12	.25	-1.80	-.60	.60	1.80	.94						

Note. Item 1 and 2 were used as the anchor and target items, respectively. The loading and threshold values in parentheses indicate that they used for the “small DIF” conditions; adjacent values were used for the “large DIF” conditions. The parameter values for the other items are not shown in the focal group because they were equal to those in the reference group.

Table 2

Population Parameters Values Used for the Simulation Study (Continued)

	Focal Group DIF on γ					
Item	λ	γ_1	γ_2	γ_3	γ_4	β
1	.90	-1	.20	1.40	2.60	.19
2	.80	-1 (-1.50)	.20 (-.30)	1.40 (.90)	2.60 (2.10)	.36
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						

3. Procedure

In each replication, a baseline model and subsequent nested models were fitted to the generated data. Two measures of overall model fit, including Jöreskog's (1971) chi-square and Bentler's (1990) CFI were obtained in each step.

The free-baseline MACS technique, described in Chapter II, tested both uniform and non-uniform DIF. The steps for testing non-uniform DIF were as follows (see Figure 1):

1. Estimate all parameters within each group, except those needed for scaling. This step provided the chi-square and CFI values for the baseline model (Model A)
2. Estimate all parameters as in Model A, except for one change. Specifically, a constraint was imposed such that the loadings of Item 2 were equal between two groups (Model B).
3. Compute the chi-square difference and the CFI difference between Model A and Model B. If the chi-square difference exceeded a critical value with 1 degree of freedom, non-uniform DIF was flagged. Similarly, if the CFI difference exceeded a critical value, this item was identified as having non-uniform DIF.

Next, Model B was used as a baseline model for testing uniform DIF. The steps were as follows (see Figure 1):

1. Estimate all free parameters in Model B, except for one; the intercepts of Item 2 were constrained to be equal between two groups (Model C).

2. Compute the chi-square difference and the CFI difference between Model B and Model C. If these values exceeded a corresponding critical value, uniform DIF was flagged.

4. Analysis

The free-baseline MACS analysis was conducted via Mplus 5.0 (Muthén & Muthén, 1998-2007), using ML estimation. Mean structure analysis, of course, is also possible via LISREL (Jöreskog & Sörbom, 1996) and EQS (Bentler, 2006). The didactic expositions of using LISREL and EQS can be found in Jöreskog and Sörbom (1996) and Bentler (1989), respectively.

Type I error rates and power were obtained through a small utility program written in FORTRAN. Type I error rate was computed as the proportion of times that the invariant target item was incorrectly flagged as having DIF (i.e., false positive). The proportion of trials in which the null hypothesis of invariance was rejected was counted based on the LR test as well as ΔCFI test. As mentioned previously, both uncorrected and Bonferroni-corrected *p* values were used as a criterion value for the LR test. The criterion values of –0.01 and –0.002 were used for the ΔCFI test. An empirical Type I error rate greater than the nominal alpha value (.05) was considered to be inflated. Power rate was computed as the proportion of times that the non-invariant target item was correctly identified as having DIF (i.e., true negative).

Finally, variance components analysis was used to examine which condition factors influenced the performance of the MACS technique. The variance components analysis is a variant of analysis of variance, which allows for the

estimation of the variation in a given dependent variable that is accounted for by a set of independent variables. In the current study, the dependent variable was Type I error rate and power, whereas the independent variables included the type of item response, scale size, similarity of sample size, latent trait distribution, type of DIF in an anchor, type of DIF in a target item, and scaling method. All the effects in the model, except for the intercept, were considered random and minimum variance quadratic unbiased estimation (MIVQUE) was used. This analysis was conducted via SAS 9.1 (SAS Institute, 2004).

CHAPTER IV: RESULTS

This chapter presents the results of the current Monte Carlo simulation study, which examined the performance of the mean and covariance structure (MACS) confirmatory factor analysis (CFA) technique for detecting differential item functioning (DIF) under various conditions. The results presented are as follows: (1) reliability, (2) model fit, (3) Type I error, and (4) power.

1. Reliability

As described in Chapter II, dichotomous responses and polytomous responses were generated for each of three combinations of sample sizes (i.e., N_f = 100, 250, 500 and N_r = 900, 750, 500 such that $N_f + N_r$ = 1,000), crossed by two combinations of latent trait distributions. The focal group had the same trait distribution as the reference group ($\xi \sim N[0,1]$), or the trait mean differed by 1 standard deviation so that the former group had a smaller trait mean ($\xi \sim N[-1,1]$). The reliability of the scale was supported for both types of response. In the dichotomous cases, Cronbach's alpha ranged from 0.91 to 0.95 [7], and it was a little higher when two groups had equal sample sizes or unequal trait means. In the polytomous cases, the alpha was equal to 0.97, regardless of the similarity of the sample sizes and the trait mean difference.

2. Model Fit

Table 3 presents average chi-square and CFI values of a baseline model for each combination of four condition factors. These fit values indicated that the

[7] Helmstadter (1964) noted that the Cronbach's alpha of 0.50 is acceptable for comparisons of two group means differing by one fourth of a standard deviation.

baseline models fit at least adequately when the scale consisted of six items. The CFI values were equal to or greater than 0.85 in the dichotomous cases and near 0.90 in the polytomous cases (see Bentler & Bonett, 1980). When the scale was longer (i.e., 12 items), however, the baseline models did not fit adequately; the CFI value ranged from .53 to .59. The chi-square value of the baseline models was substantial, rejecting the null hypothesis of exact model fit.

Table 3

Mean Model Fit Values for the Baseline Model

DIF Tested	Item response	Scale size	Trait Mean	Chi-square	CFI
Non-uniform	Dichotomous	6 items	Equal	1441.89	.85
			Unequal	1393.37	.85
		12 items	Equal	10886.30	.53
			Unequal	10377.61	.53
	Polytomous	6 items	Equal	1384.18	.87
			Unequal	1344.03	.87
		12 items	Equal	11081.77	.58
			Unequal	10743.97	.59
Uniform	Dichotomous	6 items	Equal	1424.44	.85
			Unequal	1403.10	.85
		12 items	Equal	10852.63	.53
			Unequal	10419.15	.53
	Polytomous	6 items	Equal	1380.72	.87
			Unequal	1346.02	.87
		12 items	Equal	11098.58	.58
			Unequal	10758.21	.59

Note. The degrees of freedom were 18 and 108 in the 6-item and 12-item conditions, respectively.

3. Type I Error and Power

The Monte Carlo simulation results are presented separately for Type I error and power. Then, they are visually summarized for each of three scaling methods.

3.1. Type I Error

Type I error, by all combinations of the condition factors, appears in Appendix B. As mentioned in the method chapter, Type I error was obtained separately for two test statistics and their different criterion values; uncorrected and Bonferroni-corrected *p* values of the LR test and ΔCFI values of –0.01 and –0.002. The results from using the conventional, uncorrected LR test are presented first, followed by those from using the three alternative criterion values.

3.1.1. Uncorrected *p* Value of the LR Test

When DIF was not simulated on the scale and the latent trait means were equal between groups, both marker-variable and fixed-factor scaling methods provided reasonable control for Type I error. Their Type I error rates were below or near the nominal alpha value (.05), regardless of the type of DIF being tested. Some exceptions occurred for the fixed-factor method: when uniform DIF was located among binary items with the group sizes of grater than 100 (.136 - .166). In contrast, Type I error for the effects-coded scaling method was generally inflated above the nominal alpha value in most conditions. Nevertheless, this method provided acceptable Type I error rates when used to locate uniform DIF among six ordinal items.

When large DIF was simulated on the anchor (or scale), both the marker-variable and effects-coded methods lost their control for Type I error. The Type I error rates largely exceeded the nominal alpha value in most conditions (.050 - 1). In general, the presence of DIF in the scale (but not in the target item) also inflated Type I error for the fixed-factor method. However, this method still controlled Type I error at the nominal alpha level in some conditions; when the scale consisted of six ordinal items (0 - .032) and when the scale had 12 ordinal items and one of them was contaminated by uniform DIF (.002 - .026). These Type I error results are summarized in Table 4.

Table 4

Type I error of the Uncorrected LR Test in the Equal Latent Trait Mean Condition

	Tested	Scaling method		
Anchor Type	DIF type	Marker-variable	Fixed-factor	Effects-coded
Unbiased	Non-uniform	O	O	X
	Uniform	O	O	X
DIF	Non-uniform	X	O (6 ordinal items)	X
	Uniform	X	O (6 ordinal items)	X

Note. The "O" and "X" indicate the conditions where, in general, Type I error was controlled or inflated at the nominal alpha level, respectively.

When latent trait means differed by 1 standard deviation between groups, Type I error was severely inflated in most conditions, regardless of the scaling method. In addition, the inflation was more severe when uniform DIF was tested. For example, the rate of falsely identifying the unbiased target item as having uniform DIF easily approached 1. Nevertheless, some exceptions to the inflated Type I error

occurred when non-uniform DIF was tested with polytomous responses; with an unbiased anchor and the marker-variable method (0 - .004); and with an unbiased or uniform DIF anchor and the fixed-factor method (0 - .010).

3.1.2. Alternative Criterion Values

In general, the use of an alternative criterion value (i.e., Bonferroni-corrected *p*, ΔCFI) decreased Type I error in the MACS technique.

The overall Type I error reduction due to the Bonferroni correction was the greatest in the fixed-factor scaling method (38.1% decreases), followed by the effects-coded method (19.4%) and the marker-variable method (8.0%). As a result, Type I error for the fixed-factor method was almost eliminated, except for a few conditions (e.g., when group sizes were greater than 100; when non-uniform DIF was tested in the scale including non-uniform DIF). In contrast, the Type I error rate for the marker-factor method was always inflated above the nominal alpha value when the anchor was biased. It was difficult to find consistent patterns of when the effects-coded method provided acceptable Type I error rates.

The use of ΔCFI markedly decreased Type I error in most conditions. Consequently, the Type I error rates were below the nominal alpha value unless the focal group had a sample size of 100 (fixed-factor method) and the anchor was biased (marker-variable method, effects-coded method). The reduction in Type I error was more prominent when the ΔCFI test of –0.01 was used with binary items. Using the ΔCFI test of –0.002 had positive effects on Type I error, yielding similar Type I error

results as using the Bonferroni correction. Table 5 summarizes the Type I error results from using an alternative criterion value when the anchor was biased.

Table 5

Type I error of Using an Alternative Criterion in the Biased-Anchor, Equal Latent Trait Mean Condition

Test statistic	Tested DIF type	Scaling method		
		Marker-variable	Fixed-factor	Effects-coded
Corrected *p*	Non-uniform	O (6 ordinal items)	O (6 ordinal items)	O (6 ordinal items)
	Uniform	X	O (N_f > 100)	X
ΔCFI	Non-uniform	O (ordinal items)	O (N_f > 100)	O (ordinal items)
	Uniform	X	O (N_f > 100)	X

When the trait means were not equal, Type I error reduction due to the Bonferroni correction was relatively small, compared to the case of equal trait means. The overall reduction ranged from 2.8% (fixed-factor method) to 15.0% (effects-coded method). On the other hand, using the ΔCFI test of –0.01 still almost eliminated Type I error in some conditions. For example, the Type I error rate was below the nominal alpha value when non-uniform DIF was located among ordinal items; when non-uniform DIF was located among binary items, the fixed-factor method was used, and group sizes were greater than 100; when uniform DIF was tested with the marker-variable method and an unbiased or non-uniform DIF anchor; when uniform DIF was located among 12 ordinal items using the fixed-factor method; and when uniform DIF was located among ordinal items using the effects-coded method. Using the ΔCFI test of –0.002 resulted in similar patterns of Type I error

reduction, but the reduced amounts were much smaller than using the ΔCFI test of –0.01. Table 6 summarizes the results from using an alternative criterion value when the anchor was biased and the trait means were not equal between groups.

Table 6

Type I error of Using an Alternative Criterion in the Biased-Anchor, Unequal Latent Trait Mean Condition

	Tested	Scaling method		
Test statistic	DIF type	Marker-variable	Fixed-factor	Effects-coded
Corrected *p*	Non-uniform	X	O (6 ordinal items)	X
	Uniform	X	X	X
ΔCFI	Non-uniform	O (ordinal items)	O (ordinal items)	O (ordinal items)
	Uniform	X	X	X

3.1.3. Results of Variance Components Analysis

The results of the variance components analysis are presented in Table 7. The MIVQUE estimates reported in this table reflect the amounts of variation in Type I error that are accounted for by each condition factor. In other words, the higher the estimate value, the more the corresponding condition factor contributed to Type I error in the MACS technique.

In general, the results were somewhat similar across four different criterion values, which were independently used to obtain the Type I error rate. That is, the most influential single factors were commonly similarity of latent trait means, type of item response, and type of anchor.

When either the uncorrected or Bonferroni-corrected LR test was used, the most important factor was the similarity of latent trait means (0.065 and 0.058 for uncorrected and corrected *p* values, respectively). It was followed by the two-way interaction between type of DIF and type of item response. The MIVQUE estimates were 0.032 and 0.038 for uncorrected and corrected *p* values, respectively. Type I error was also strongly influenced by another two-way interaction between scaling method and type of anchor (0.024 and 0.026 for uncorrected and corrected *p* values, respectively).

When the ΔCFI test of –0.01 was used, the six-way interaction, scaling method × similarity of latent trait means × type of DIF × type of item response × scale size × type of anchor, contributed the most to the Type I error variance (0.022). For the ΔCFI test of –0.002, the highest interaction term (i.e., seven-way) accounted for the most variance (0.054). Then, it was followed by the three-way interaction, scaling method × latent distribution × type of DIF (0.032), and the two-way interaction between scaling method and type of anchor (0.025).

Table 7

Results of the Variance Components Analysis for Type I Error

	Criterion value			
	LR test		ΔCFI test	
Factor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Scaling method (S)	0	0	0	0
Latent trait mean (L)	0.065	0.058	0.004	0.015
Type of DIF (D)	0	0	0	0
Item response (I)	0.012	0.010	0.002	0.016
Scale size (T)	0.002	0.001	0	0
Sample size (N)	0	0	0	0
Type of Anchor (A)	0.008	0.008	0.002	0.009
S*L	0.020	0.020	0	0.011
S*D	0.009	0.004	0	0
S*I	0.002	0	0.009	0.002
S*T	0	0	0	0
S*N	0	0	0.002	0
S*A	0.024	0.026	0.006	0.025
L*D	0.031	0.024	0	0
L*I	0	0	0.002	0
L*T	0	0	0	0.001
L*N	0.002	0.005	0	0.001
L*A	0	0	0.001	0
D*I	0.032	0.038	0.002	0.019
D*T	0.002	0.002	0.001	0.001
D*N	0	0	0.003	0.002
D*A	0	0	0	0.002
I*T	0	0	0.001	0
I*N	0.003	0.005	0.021	0.012
I*A	0.003	0.004	0.003	0.004
T*N	0	0	0.001	0.002
T*A	0.002	0.002	0	0.001
N*A	0	0.001	0.001	0
S*L*D	0	0	0	0.032
S*L*I	0	0.007	0	0.001
S*L*T	0.001	0.002	0.007	0.007

Note. The variance components estimates should theoretically be positive because they represent the variance of a random variable. Under the assumption that the fitted random effects model was appropriate for the data, the negative estimate value, if present, was considered as zero following common practice.

Table 7

Results of the Variance Components Analysis for Type I Error (Continued)

	Criterion value			
	LR test		ΔCFI test	
Factor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
S*L*N	0	0	0	0
S*L*A	0	0.002	0	0.003
S*D*I	0	0	0	0.001
S*D*T	0.001	0.001	0	0.003
S*D*N	0	0	0	0
S*D*A	0	0	0.001	0.008
S*I*T	0.001	0.001	0	0.001
S*I*N	0.006	0.009	0.002	0.012
S*I*A	0	0.001	0	0
S*T*N	0	0	0.004	0
S*T*A	0.003	0.004	0.002	0.004
S*N*A	0	0	0.002	0.003
L*D*I	0.007	0.009	0	0.009
L*D*T	0.002	0.002	0.003	0.001
L*D*N	0.001	0.001	0.002	0.002
L*D*A	0.026	0.021	0	0
L*I*T	0.002	0.003	0	0.001
L*I*N	0.001	0	0	0
L*I*A	0	0	0	0
L*T*N	0	0	0.002	0
L*T*A	0.002	0.003	0	0
L*N*A	0	0	0	0.005
D*I*T	0.001	0.002	0	0.001
D*I*N	0.001	0	0	0
D*I*A	0	0	0	0.001
D*T*N	0	0	0	0
D*T*A	0.009	0.008	0	0
D*N*A	0	0	0.001	0
I*T*N	0.001	0	0	0
I*T*A	0	0	0	0
I*N*A	0	0	0	0
T*N*A	0	0	0	0.001
S*L*D*I	0.024	0.014	0.005	0
S*L*D*T	0	0	0.006	0
S*L*D*N	0.001	0.002	0	0
S*L*D*A	0.015	0.005	0.001	0

Table 7

Results of Variance Components Analysis for Type I Error (Continued)

	Criterion value			
	LR test		ΔCFI test	
Factor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
S*L*I*T	0	0	0.003	0
S*L*I*N	0	0.002	0.003	0.004
S*L*I*A	0	0	0.003	0
S*L*T*N	0.001	0.001	0	0
S*L*T*A	0.001	0.001	0.001	0
S*L*N*A	0	0	0	0
S*D*I*T	0	0	0.004	0
S*D*I*N	0.002	0.003	0.001	0.001
S*D*I*A	0	0	0.010	0.013
S*D*T*N	0	0.001	0.001	0
S*D*T*A	0	0	0.005	0
S*D*N*A	0.003	0.004	0.001	0
S*I*T*N	0.001	0.003	0.002	0.003
S*I*T*A	0.001	0	0.006	0.001
S*I*N*A	0.002	0.002	0	0.002
S*T*N*A	0	0.002	0	0
L*D*I*T	0	0	0.002	0
L*D*I*N	0	0	0.002	0
L*D*I*A	0.003	0.003	0.003	0.004
L*D*T*N	0	0	0	0.002
L*D*T*A	0	0	0.002	0.002
L*D*N*A	0	0.002	0.001	0
L*I*T*N	0	0	0.001	0
L*I*T*A	0	0	0.005	0.004
L*I*N*A	0.001	0.002	0.003	0
L*T*N*A	0	0	0	0
D*I*T*N	0	0	0.003	0.001
D*I*T*A	0.003	0	0.004	0.001
D*I*N*A	0	0.001	0	0.001
D*T*N*A	0.002	0.003	0	0
I*T*N*A	0.001	0.002	0.004	0.002
S*L*D*I*T	0.003	0.004	0	0.016
S*L*D*I*N	0	0	0	0.002
S*L*D*I*A	0.020	0.022	0	0.014
S*L*D*T*N	0	0	0.021	0.007
S*L*D*T*A	0.001	0.001	0	0.004

Table 7

Results of Variance Components Analysis for Type I Error (Continued)

	Criterion value			
	LR test		ΔCFI test	
Factor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
S*L*D*N*A	0.002	0.003	0.001	0.009
S*L*I*T*N	0	0	0	0
S*L*I*T*A	0	0	0	0.005
S*L*I*N*A	0.001	0.002	0.003	0.008
S*L*T*N*A	0	0	0.004	0.008
S*D*I*T*N	0	0	0	0.002
S*D*I*T*A	0.002	0.004	0	0
S*D*I*N*A	0	0	0.003	0.006
S*D*T*N*A	0	0	0.001	0.014
S*I*T*N*A	0	0	0.001	0.002
L*D*I*T*N	0.002	0.002	0	0.003
L*D*I*T*A	0.003	0.003	0	0
L*D*I*N*A	0	0	0	0.003
L*D*T*N*A	0.001	0	0	0.002
L*I*T*N*A	0	0	0	0
D*I*T*N*A	0	0	0	0.002
S*L*D*I*T*N	0.002	0.004	0	0
S*L*D*I*T*A	0.002	0.004	0.022	0
S*L*D*I*N*A	0	0	0	0
S*L*D*T*N*A	0	0	0	0
S*L*I*T*N*A	0.003	0.003	0	0
S*D*I*T*N*A	0.002	0.003	0.002	0
L*D*I*T*N*A	0.002	0.003	0.002	0
S*L*D*I*T*N*A	0.006	0.009	0.018	0.054

3.2. Power

Power, by all combinations of the condition factors, appears in Appendix C. First of all, descriptive analysis was used to investigate which condition factors considerably affected power. The mean power for each of four different criterion values is presented in Table 8, crossed by five condition factors. These five condition factors were those found the most influential single factors in the previous variance components analysis for Type I error.

On average, power was higher under these conditions; when DIF is large (.537) rather than small (.425); when latent trait means were unequal between groups (.531) rather than equal (.431); when item responses were dichotomous (.566) rather than polytomous (.397); when uniform DIF was tested (.625) rather than non-uniform DIF (.337); and the anchor was non-invariant (.499) rather than invariant (.463). In addition, using a conventional, uncorrected *p* value of the LR test provided the highest power (.695). Then, this was followed by using Bonferroni correction (.596), a ΔCFI value of –0.002 (.383), and a ΔCFI value of –0.01 (.179). The difference in average power was negligible between invariant anchor and non-invariant anchor across the four different criterion values (.013 - .059).

Table 8

Mean Power Rate by Five Condition Factors.

Factor					Criterion value			
					LR test		ΔCFI test	
E	L	D	I	A	Uncorrected *p*	Corrected *p*	–0.01	–0.002
E1	L1	D1	I1	A1	.479	.355	.155	.214
				A2	.594	.531	.251	.445
			I2	A1	.254	.091	.000	.003
				A2	.312	.160	.020	.036
		D2	I1	A1	.796	.677	.112	.264
				A2	.834	.734	.117	.407
			I2	A1	.785	.692	.000	.242
				A2	.838	.729	.048	.352
	L2	D1	I1	A1	.790	.662	.177	.358
				A2	.798	.682	.352	.490
			I2	A1	.273	.114	.000	.002
				A2	.373	.269	.023	.084
		D2	I1	A1	.643	.612	.345	.559
				A2	.933	.901	.428	.740
			I2	A1	.906	.757	.202	.410
				A2	.901	.871	.286	.742
E2	L1	D1	I1	A1	.738	.640	.164	.351
				A2	.730	.644	.204	.430
			I2	A1	.596	.446	.060	.149
				A2	.455	.304	.030	.095
		D2	I1	A1	.903	.875	.292	.771
				A2	.800	.747	.317	.650
			I2	A1	.865	.833	.266	.750
				A2	.657	.600	.195	.531
	L2	D1	I1	A1	.680	.615	.189	.454
				A2	.796	.735	.313	.552
			I2	A1	.541	.392	.026	.087
				A2	.441	.288	.016	.085
		D2	I1	A1	.881	.790	.546	.730
				A2	.951	.921	.570	.805
			I2	A1	.993	.981	.339	.781
				A2	.916	.841	.431	.619

Note. E1 = small DIF, E2 = large DIF; L1 = equal latent trait means, L2 = unequal latent trait means, D1 = non-uniform DIF, D2 = uniform DIF; I1 = dichotomous item response, I2 = polytomous item response, A1 = invariant anchor, A2 = non-invariant anchor.

3.2.1. Uncorrected *p* Value of the LR Test

When the latent trait means were equal, power for detecting uniform DIF was adequate for both marker-variable and fixed-factor scaling methods. That is, power was always greater than .80, regardless of the type of item response and DIF size. Only exceptions occurred when large uniform DIF was tested with the marker-variable method and a uniform DIF anchor (.012 - .557). In contrast, the effects-coded method provided adequate power in some conditions, but it was difficult to find consistent patterns.

When used to detect large non-uniform DIF under the equal trait means, the marker-variable method provided adequate power with the 12-item scale including an invariant anchor (.804 - 1). However, with the 6-item scale containing an invariant anchor, power was marginal (.167 - .744). Under the equal trait means, power for detecting large non-uniform DIF was adequate for the fixed-factor method only in a few conditions; when group sizes were greater than 100, the scale consisted of six binary items, and this scale included another uniform DIF item or DIF-free items. When the item responses were polytomous, however, power for detecting non-uniform DIF was always less than .80 for this scaling method (.386 - .564), regardless of the DIF size. Under the same condition (e.g., equal trait means, testing non-uniform DIF), the effects-coded method provided adequate power when the scale included an invariant anchor (.835 - 1).

In general, unequal trait means did increase power in the MACS technique. In addition, the increase was the greatest when uniform DIF was tested with the effects-

coded method. Consequently, this scaling method provided adequate power for detecting large uniform DIF unless the scale included another uniform DIF item (.800 - 1). Interestingly, negligible but small decrease in overall power for detecting small DIF was observed for the marker-variable method (6.2%).

3.2.2. Alternative Criterion Values

Generally, the use of an alternative criterion value reduced power in the MACS technique. Nevertheless, for both marker-variable and fixed-factor scaling methods, Bonferroni correction did not adversely affect power for detecting large uniform DIF. For detecting small uniform DIF, using the Bonferroni-corrected LR test provided adequate power when group sizes were greater than 100.

The ΔCFI test was somewhat stringent in the sense that power for any scaling method was not adequate in most conditions. For example, power of the ΔCFI test of –0.01 could not reach .80 except for only a few conditions. Nevertheless, power for detecting large uniform DIF was greater than .80 when the ΔCFI value of –0.002 was used with the fixed-factor scaling method and the group sizes of greater than 100. For the marker-variable method, power for detecting large uniform was adequate when the ΔCFI value of –0.002 was used, group sizes were greater than 100, and the anchor was unbiased or non-uniform DIF item. For the effects-coded method, power was less than .80 in nearly all conditions.

Unequal trait means improved power in the MACS technique especially when uniform DIF was tested. Consequently, with the ΔCFI value of –0.002 and the fixed-factor scaling method, power for detecting uniform DIF approached 1 in almost all

conditions. Although power also increased for the other two scaling methods, any consistent pattern could not be observed.

3.2.3. Results of Variance Components Analysis

Table 9 presents the results of the variance components analysis. In general, the most influential single factors were commonly similarity of latent trait means and type of item response across four different criterion values.

For the LR test, the most important factor was the two-way interaction between similarity of latent trait means and type of item response (0.024 for uncorrected *p* value) and the type of item response (0.034 for corrected *p* value). For the ΔCFI test of –0.01, the five-way interaction, scaling method × type of DIF × type of item response × scale size × type of anchor, contributed the most to the power variance (0.018). For the ΔCFI test of –0.002, the two-way interaction between sample sizes and type of anchor accounted for the most variance (0.028).

These results were somewhat different from those obtained for Type I error. For example, although the most influential single factors were in common between Type I error and power, power was not strongly influenced by the two-way interaction between scaling method and type of anchor (0.001 - 0.008).

The Type I error and power results are visually summarized in Tables 10 through 13. The shaded areas in these tables indicate the cases where the MACS technique performed well to test at least small DIF, in terms of Type I error and power. In other words, if the Type I error rate was less than .05 and power was greater than .80, the corresponding conditions were shaded in this table.

Table 9

Results of the Variance Components Analysis for Power

	Criterion value			
	LR test		ΔCFI test	
Factor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Scaling method (S)	0	0	0	0
Latent trait mean (L)	0.021	0.020	0.004	0.004
Type of DIF (D)	0.008	0.009	0	0.016
Item response (I)	0.003	0.001	0	0
Scale size (T)	0	0	0	0
Sample size (N)	0	0	0	0
Type of Anchor (A)	0.002	0	0	0
S*L	0.021	0.025	0	0.021
S*D	0.001	0.002	0	0
S*I	0	0	0	0.003
S*T	0	0	0.009	0.004
S*N	0	0	0	0.003
S*A	0	0	0.002	0
L*D	0.001	0.004	0.001	0.007
L*I	0	0	0	0
L*T	0	0	0	0.001
L*N	0	0.001	0.002	0.002
L*A	0	0	0	0.001
D*I	0.024	0.034	0.002	0.028
D*T	0	0	0.005	0.009
D*N	0	0.001	0.008	0.006
D*A	0.001	0.003	0	0.001
I*T	0	0	0.001	0
I*N	0.003	0.004	0.018	0.008
I*A	0.003	0.005	0.001	0.004
T*N	0	0	0.003	0.002
T*A	0	0	0	0.001
N*A	0	0.001	0	0
S*L*D	0	0.001	0.011	0.014
S*L*I	0.008	0.008	0.003	0
S*L*T	0.002	0.003	0	0

Note. The dependent variable was power averaged for small and large DIF conditions. Under the assumption that the fitted random effects model was appropriate for the data, the negative estimate value, if present, was considered as zero following common practice.

Table 9

Results of the Variance Components Analysis for Power (Continued)

	Criterion value			
	LR test		ΔCFI test	
Factor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
S*L*N	0	0	0	0
S*L*A	0.006	0.006	0	0
S*D*I	0	0	0.003	0.007
S*D*T	0.004	0.006	0.006	0
S*D*N	0	0	0	0
S*D*A	0.001	0	0	0
S*I*T	0.001	0.001	0	0
S*I*N	0.005	0.008	0.001	0.013
S*I*A	0.002	0.002	0.002	0.005
S*T*N	0	0	0.002	0
S*T*A	0	0	0	0
S*N*A	0	0	0	0
L*D*I	0.005	0.004	0	0
L*D*T	0.001	0.001	0.001	0
L*D*N	0	0	0	0
L*D*A	0	0	0	0
L*I*T	0.001	0.001	0	0.001
L*I*N	0	0	0	0
L*I*A	0	0	0.001	0
L*T*N	0	0	0	0
L*T*A	0	0	0	0
L*N*A	0.001	0	0	0
D*I*T	0.004	0.004	0	0.001
D*I*N	0	0	0	0
D*I*A	0	0	0.001	0
D*T*N	0	0	0	0
D*T*A	0.001	0	0	0
D*N*A	0	0	0.001	0.002
I*T*N	0	0	0	0
I*T*A	0	0	0	0
I*N*A	0	0	0	0.001
T*N*A	0	0	0	0
S*L*D*I	0.004	0	0	0
S*L*D*T	0	0	0	0.005
S*L*D*N	0	0.003	0.007	0.001
S*L*D*A	0	0	0	0

Table 9

Results of Variance Components Analysis for Power (Continued)

	Criterion value			
	LR test		ΔCFI test	
Factor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
S*L*I*T	0	0	0	0
S*L*I*N	0.001	0.001	0.002	0
S*L*I*A	0	0.001	0	0
S*L*T*N	0	0	0.002	0.001
S*L*T*A	0	0	0.001	0
S*L*N*A	0	0.001	0.005	0.004
S*D*I*T	0	0	0	0.002
S*D*I*N	0	0.001	0.002	0.001
S*D*I*A	0	0.001	0.001	0.001
S*D*T*N	0	0.001	0	0.001
S*D*T*A	0	0.001	0.003	0.003
S*D*N*A	0.001	0.001	0	0.001
S*I*T*N	0.001	0.001	0	0.001
S*I*T*A	0.001	0.001	0	0.002
S*I*N*A	0	0	0.001	0
S*T*N*A	0	0.001	0.001	0.002
L*D*I*T	0.005	0.007	0	0.007
L*D*I*N	0.001	0.001	0	0.002
L*D*I*A	0	0	0.001	0.002
L*D*T*N	0	0	0	0.001
L*D*T*A	0.001	0.001	0.002	0.001
L*D*N*A	0.001	0.001	0.001	0
L*I*T*N	0	0.001	0.002	0.001
L*I*T*A	0	0	0.001	0.002
L*I*N*A	0.001	0.001	0.001	0.001
L*T*N*A	0	0	0.002	0
D*I*T*N	0.001	0.002	0	0
D*I*T*A	0	0	0	0
D*I*N*A	0.001	0.002	0.003	0.002
D*T*N*A	0.004	0.008	0	0.004
I*T*N*A	0.001	0	0	0
S*L*D*I*T	0.004	0.005	0.004	0.002
S*L*D*I*N	0	0	0	0.002
S*L*D*I*A	0	0	0	0
S*L*D*T*N	0	0	0	0
S*L*D*T*A	0.001	0.001	0.002	0

Table 9

Results of Variance Components Analysis for Power (Continued)

	Criterion value			
	LR test		ΔCFI test	
Factor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
S*L*D*N*A	0	0	0	0
S*L*I*T*N	0	0	0	0
S*L*I*T*A	0	0	0	0.003
S*L*I*N*A	0	0	0	0
S*L*T*N*A	0	0	0	0
S*D*I*T*N	0	0	0	0
S*D*I*T*A	0.002	0.001	0	0
S*D*I*N*A	0	0	0.002	0.001
S*D*T*N*A	0	0	0	0
S*I*T*N*A	0	0	0	0
L*D*I*T*N	0	0	0	0
L*D*I*T*A	0	0	0	0
L*D*I*N*A	0	0	0	0
L*D*T*N*A	0	0	0	0
L*I*T*N*A	0	0	0	0
D*I*T*N*A	0	0	0	0.003
S*L*D*I*T*N	0.001	0.002	0.009	0.003
S*L*D*I*T*A	0.001	0.001	0.002	0.001
S*L*D*I*N*A	0.001	0.001	0.001	0.001
S*L*D*T*N*A	0.001	0	0.001	0.001
S*L*I*T*N*A	0.003	0.005	0.004	0.006
S*D*I*T*N*A	0.002	0.002	0.002	0.002
L*D*I*T*N*A	0.004	0.006	0.003	0.001
S*L*D*I*T*N*A	0.001	0.002	0.002	0.004

Table 10

Testing Non-Uniform DIF when Latent Trait Means are Equal

Factor				Scaling method											
				Marker-variable				Fixed-factor				Effects-coded			
I	T	N	A	a	b	c	d	a	b	c	d	a	b	c	d
I1	T1	N1	A1												
			A2												
			A3												
		N2	A1												
			A2												
			A3												
		N3	A1												
			A2												
			A3												
	T2	N1	A1												
			A2												
			A3												
		N2	A1												
			A2												
			A3												
		N3	A1												
			A2												
			A3												
I1	T1	N1	A1												
			A2												
			A3												
		N2	A1												
			A2												
			A3												
		N3	A1												
			A2												
			A3												
	T2	N1	A1												
			A2												
			A3												
		N2	A1												
			A2												
			A3												
		N3	A1												
			A2												
			A3												

Note. The factor I is defined in Table 5. T1 = 6-item scale, T2 = 12-item scale; N1 = 100/900 sample sizes, N2 = 250/750 sample sizes, N3 = 500/500 sample sizes; A1 = invariant anchor, A2 = non-uniform DIF anchor, A3 = uniform DIF anchor; a = uncorrected LR test, b = Bonferroni-corrected LR test, c = ΔCFI test of –0.01, d = ΔCFI test of –0.002.

Table 11

Testing Uniform DIF when Latent Trait Means are Equal

Factor				Scaling method											
				Marker-variable				Fixed-factor				Effects-coded			
I	T	N	A	a	b	c	d	a	b	c	d	a	b	c	d
I1	T1	N1	A1	■				■	■						
			A2	■											
			A3					■							
		N2	A1	■	■				■						
			A2		■				■						
			A3						■						
		N3	A1	■	■				■		■				
			A2	■	■				■		■				
			A3						■		■				
	T2	N1	A1	■				■							
			A2												
			A3												
		N2	A1	■					■						
			A2		■				■						
			A3						■						
		N3	A1	■	■				■						
			A2	■	■				■						
			A3						■						
I1	T1	N1	A1	■	■			■							
			A2	■	■			■							
			A3					■							
		N2	A1	■	■		■	■	■		■				
			A2		■		■	■	■		■				
			A3					■	■		■				
		N3	A1	■	■		■	■	■		■	■			
			A2	■	■		■	■	■		■	■			
			A3					■	■		■				
	T2	N1	A1	■	■			■							
			A2	■	■										
			A3					■							
		N2	A1	■	■			■	■						
			A2		■			■	■						
			A3					■	■						
		N3	A1	■	■			■	■						
			A2	■	■			■	■						
			A3					■	■						

Table 12

Testing Non-Uniform DIF when Latent Trait Means are Unequal

Factor				Scaling method											
				Marker-variable				Fixed-factor				Effects-coded			
I	T	N	A	a	b	c	d	a	b	c	d	a	b	c	d
I2	T1	N1	A1												
			A2												
			A3												
		N2	A1												
			A2												
			A3												
		N3	A1												
			A2												
			A3												
	T2	N1	A1												
			A2												
			A3												
		N2	A1												
			A2												
			A3												
		N3	A1												
			A2												
			A3												
I2	T1	N1	A1												
			A2												
			A3												
		N2	A1												
			A2												
			A3												
		N3	A1												
			A2												
			A3												
	T2	N1	A1												
			A2												
			A3												
		N2	A1												
			A2												
			A3												
		N3	A1												
			A2												
			A3												

Table 13

Testing Uniform DIF when Latent Trait Means are Unequal

Factor				Scaling method											
				Marker-variable				Fixed-factor				Effects-coded			
I	T	N	A	a	b	c	d	a	b	c	d	a	b	c	d
I2	T1	N1	A1												
			A2												
			A3												
		N2	A1												
			A2												
			A3												
		N3	A1												
			A2												
			A3												
	T2	N1	A1												
			A2												
			A3												
		N2	A1												
			A2												
			A3												
		N3	A1												
			A2												
			A3												
I2	T1	N1	A1												
			A2												
			A3												
		N2	A1												
			A2												
			A3												
		N3	A1												
			A2												
			A3												
	T2	N1	A1												
			A2												
			A3												
		N2	A1												
			A2												
			A3												
		N3	A1												
			A2												
			A3												

CHAPTER V: SUMMARY AND DISCUSSION

This chapter discusses the study findings in the context of educational and psychological assessment in the following order: (1) summary of the study, (2) summary of the study findings, (3) confirmed research hypotheses, (4) discussions and implications, (5) limitations and future directions, and (6) novel contribution and conclusion.

1. Summary of the Study

Given that mean and covariance structure (MACS) confirmatory factor analysis (CFA) has enjoyed increasing attention in the differential item functioning (DIF) literature, the primary purpose of this dissertation was to evaluate the performance of MACS analysis for DIF detection. Although different scaling methods can lead to different conclusions about DIF, this issue had not been fully examined.

Accordingly, this dissertation presents an empirical study that examined the Type I error and power of the MACS technique by means of Monte Carlo simulation. The manipulated condition factors included type of item response, scale size, similarity of sample sizes, type of DIF, amount of DIF, similarity of latent trait distributions, type of anchor, test statistic and its criterion value, and scaling method.

2. Summary of the Study Findings

2.1. Type I Error

It appeared that overall, three different scaling methods provide different Type I error rates. Specifically, when the scale included only DIF-free items (equivalently

when the anchor was invariant), the latent trait means were equal, and the conventional, uncorrected LR test was used, both marker-variable and fixed-factor scaling methods provided reasonable control for Type I error at the nominal alpha level. On the other hand, effects-coded scaling method falsely indicated the presence of DIF at the rates well above the nominal alpha value.

Additionally, the current study found that the MACS technique has inflated Type I error associated with the presence of DIF in the scale (or anchor). Indeed, when the scale (or anchor) was contaminated by DIF, only the fixed-factor method provided reasonable control for Type I error under several conditions. The inflation in Type I error was especially severe when non-uniform DIF was located among binary items. These findings were supported by the subsequent variance components analysis; in cases where the uncorrected LR test was used, the four conditions factors including type of anchor, type of scaling method, type of item response, and type of DIF, individually and collectively, contributed the most to the variance of Type I error.

In general, the use of an alternative criterion value (i.e., Bonferroni-corrected *p*, ΔCFI) reduces Type I error in the MACS technique. The reduction was pronounced when the fixed-factor method was used for scaling. Indeed, when groups had comparable large sample sizes (i.e., $N_F = 250/N_R = 750$, $N_F = 500/N_R = 500$), Type I error for the fixed-factor method was almost eliminated, regardless of the type of item response, type of DIF, and scale size. On the other hand, Type I error rates for the

other scaling methods were still inflated above the nominal alpha value when the anchor was biased.

Finally, it was observed that Type I error in the MACS technique is much influenced by group differences in the latent mean. When the latent means differed by 1 standard deviation, Type I error was severely inflated in most conditions. This finding corresponds to the subsequent analysis showing that similarity of latent trait distributions was the most influential condition factor for Type I error. Nevertheless, the use of an alternative criterion value, especially ΔCFI, positively affected Type I error, providing acceptable Type I error rates for locating non-uniform DIF among ordinal items.

2.2. Power

The current simulation results indicate that, on average, the MACS technique provides higher power when large uniform DIF is detected among binary items, the anchor was non-invariant, and the latent trait means are unequal between groups.

In cases where the conventional, uncorrected LR test was used and the trait means were equal, power for detecting non-uniform DIF was not adequate for any scaling method in most conditions. However, fixed-factor scaling method provided adequate power for detecting uniform DIF. The same power for the marker-variable method was adequate only when the anchor was unbiased.

In addition, the use of an alternative criterion value was found to considerably reduce power in the MACS technique. The reduction was prominent when the ΔCFI test of –0.01 was used; it provided power less than .80 in almost all conditions. In

contrast, Bonferroni correction did not adversely affect power when large uniform DIF was tested with the group sizes of greater than 100. Finally, it was found that unequal trait means generally increase power in the MACS technique.

3. Supported Study Hypotheses

3.1. Hypothesis 1

In terms of Type I error and power, the current study found that statistically equivalent scaling methods provide different outcomes when measurement invariance is evaluated at the item level. Under favorable circumstances (i.e., comparable large group sizes, equal latent trait distributions, Bonferroni-corrected LR test, and no DIF items other than the target item), both marker-variable and fixed-factor scaling methods tested uniform DIF reasonably well. Under the same circumstances, however, effects-coded scaling method performed well only in some conditions. These findings partially support Hypothesis 1 of this dissertation; if the item parameters used for scaling are truly invariant across groups, the performance of the MACS technique will be equivalent, regardless of the scaling method.

3.2. Hypothesis 2

Under less than favorable circumstances (e.g., the scale included another DIF item or a DIF anchor), Type I error in the MACS technique was inflated above the nominal alpha level, most notably when marker-variable or effects-coded method was used for scaling. More specifically, the inflation was prominent if the anchor was biased by the same type of DIF being tested for the target item. In contrast, the fixed-factor method effectively tested uniform DIF when group sizes were greater than 100.

These findings partially support Hypothesis 2; if the item parameters used for scaling are not truly invariant across groups, the performance of the MACS technique will depend on the choice of scaling method.

4. Discussions and Implications

Measurement equivalence is a critical concern in psychological and educational research because it is often required for meaningful group comparisons. Although researchers in these fields have applied different methodologies to this issue, confirmatory factor analysis (CFA), or more broadly structural equation modeling, has offered an integrative framework in which measurement equivalence is evaluated at the item level, at the scale level, or at both. Indeed, CFA can reflect the item response theory (IRT) concept of differential functioning, while providing a variety of options (e.g., multiple latent trait variables, more than two groups, categorical or continuous covariates). The empirical results of this study bring up some methodological issues and recommendations to be considered when a researcher conducts DIF analysis using CFA.

The current simulation results appear to support the utility of the MACS technique in some circumstances and not in others. For example, poor performance was uniformly observed when groups had truly different latent trait means. This finding is, in part, consistent with the previous simulation study conducted by González-Romá et al (2006).[8] They showed that, if the trait means differ by 1

[8] They used the constrained-baseline strategy and the modification index test. For scaling, they constrained the loading and intercept of an anchor between two groups and then also constrained the latent mean only for the reference group.

standard deviation, the MACS technique controls its Type I error only when sample sizes are equal between groups. It is reasonable to argue that unequal trait means should be more concerned for a particular scaling method that constrains the trait parameters between groups (i.e., fixed-factor method). Indeed, Cheung and Rensvold (1999) noted that if trait parameters are constrained across groups when they are not actually equal, biased invariance conclusions can occur. In the current study, however, the inflation in Type I error was uniformly observed and its pattern was similar across three different scaling methods. Thus, there exists a certain risk of identifying an invariant item as a DIF item when the trait means are truly different across groups.

As mentioned previously, if present, DIF can be either item impact or item bias depending on the source of DIF (Camilli & Shepard, 1994; Zumbo, 1999). When groups truly differ in the latent trait being measured, different responses on the same scale will be observed across the groups. In this situation, item parameter values estimated from the observed responses accurately reflect true group differences in the trait (i.e., item impact). Thus, even when the population item parameter values were set to be invariant by design in the current simulation study, the estimated parameter values were very likely non-invariant between groups, (accurately) reflecting the simulated group differences in the trait mean. Taken together, the observed inflation in Type I error under the unequal latent means, in fact, might indicate high power for detecting item impact.[9] From this viewpoint, Type I error can be referred to as the

[9] If it could be reported, the power for detecting uniform DIF due to item impact appears to easily approach 1, with polytomous responses and 1 standard deviation difference in the trait mean.

probability of falsely detecting "item bias" if, and only if, the trait means are comparable across groups.

Given the necessity of the equivalent trait means for detecting item bias, a circular problem exists; (a) trait means should be comparable across groups, (b) similarity of the trait means cannot be confirmed without estimating them, (c) estimation of the trait means will be inaccurate with the presence of item bias, (d) locating item bias in the scale depends on the equality/inequality of the trait means, which brings the process back to the starting point. Accordingly, Zumbo (1999) suggests a need for a post-hoc practice, in which the "biasedness" of the flagged DIF items is determined through a series of empirical assessments and content analyses. Because detecting item impact is beyond the scope of the current study and not of general interest in the DIF literature, the following discussions are limited to the cases where a researcher wants to locate item bias in the scale.

An important issue in the measurement literature is the presence of bias in the anchor set. It has been repeatedly observed that a biased anchor set adversely affects invariance testing (Cheung & Rensvold, 1999; Finch, 2005; Navas-Ara & Gómez-Benito, 2002; Stark et al., 2006). The current study suggests a possibility that ameliorates this problem. That is, when used with the Bonferroni correction and the group sizes of greater than 100, the fixed-factor scaling method almost eliminated Type I error while maintaining adequate power for detecting (uniform) DIF. Although the ΔCFI test of –0.01 also appeared to provide reasonable control for Type I error, power was not adequate in most conditions. Similarly, French and Finch (2006) noted

that, despite the fact that the ΔCFI test of –0.01 has comparable power to the LR test (at .01 alpha level) for testing the scale-level loading invariance (i.e., metric invariance) in some conditions, this criterion rarely performs as well for testing a single loading. In addition, the ΔCFI test of –0.002 could not provide enough power to detect small uniform DIF in most conditions (see also French & Finch, 2006). Taken together, the conventional marker-variable scaling method will be suitable for testing DIF only if a designated anchor or anchor set is readily available. Otherwise, with the Bonferroni-corrected LR test and the comparable large sample sizes, the fixed-factor scaling method should be recommended for testing (at least uniform) DIF.

Combining the previous and current simulation results, a general procedural guideline for evaluating measurement invariance is suggested here. This testing procedure consists of three stages. In the first stage, omnibus metric invariance of a scale (i.e., metric invariance) is tested (see *2.5.3. Testing Procedure* in Chapter II). If metric invariance holds, then omnibus scalar invariance is evaluated in the next stage. The ΔCFI test of –0.01 (or –0.002 for high-stakes testing environments) is recommended for testing the scale-level invariance hypotheses. If it is appropriate to use maximum likelihood (ML) estimation, the conventional, Bonferroni-corrected LR test will be a comparable or better choice (see French & Finch, 2006). Because the scaling method generally does not affect the conclusions about omnibus invariance, any scaling method is applicable for testing the metric and scalar invariance.

If metric invariance is rejected, detecting item(s) having non-uniform DIF occurs within the first stage. The free-baseline MACS technique is used to examine

each item individually, using the fixed-factor scaling method and the Bonferroni-corrected LR test.[10] Then, partial scalar invariance is evaluated by implementing a condition, in which the loading and intercept parameters are constrained across groups only for the loading-invariant items.

If scalar or partial scalar invariance is rejected, the free-baseline MACS technique is used to test uniform DIF within the second stage. It is recommended that only the loading-invariant items are evaluated for uniform DIF one at a time. The fixed-factor scaling method and the Bonferroni-corrected LR test are also recommended.

Last, after locating DIF item(s), further invariance tests (e.g., unique factor variance, factor covariance, factor variance) may continue, while using the effects-coded scaling method. The baseline models should maintain the constraints of the supported partial metric and scalar invariance. As noted previously (see also Little et al., 2005), the effects-coded method provides a couple of preferable features when used to estimate the latent trait parameters in the multiple-group case. For example, the trait variable has a scale that is optimally weighted by all of its indicators. Thus, this method provides more accurate trait estimates than the fixed-factor method in which the scale is defined by a single, arbitrarily chosen anchor. Additionally, when invariance constraints are placed on the loadings and the intercepts, the effects-coded method provides the scale of the trait variable within each group, which is not the

[10] Note that these tests possibly increase Type II error when groups have truly different trait means. In other words, there exists a certain risk of falsely identifying a non-uniform DIF item as DIF-free. If present, this Type II error may adversely affect the subsequent uniform DIF tests.

case with the fixed-factor method.

5. Limitations and Future Directions

Although a number of important findings emerged in the current study, there are several weaknesses that require readers to interpret the results with caution. First, one must assess the validity of the current results because variations in conducting the MACS analysis, such as estimation method and computer software, may affect the trustworthiness of the model estimates. In this study, items were designed to have either two (binary) or five (ordinal) response categories, reflecting typical psychological or cognitive testing environments. Thus, the current study should have taken non-normality of the data into account more appropriately. The ML estimation technique, which was used in this study, assumes normality of the observed variables. When item responses are categorical, however, ML estimation can lead to erroneous invariance detection (Lubke & Muthén, 2004). Nevertheless, the validity of the current results is defendable to some extent. That is, even in the worst scenario (e.g., N_F = 250/N_R = 750), the observed polytomous responses were found to approximate a normal distribution.[11] With regard to the highly non-normal (dichotomous) responses, the use of a test statistic and/or an estimation method that are robust to the non-normality problem would have provided more reliable outcomes. For example, the Satorra-Bentler chi-square (SB chi-square; Satorra & Bentler, 1988) incorporates a scale correction to the chi-square, taking into account hypothesized model and

[11] The median skewness and kurtosis were 0.06 and -0.08, respectively. But, the responses on Item 12 were moderately non-normal for both focal and reference groups, with a skewness ranging from -3.42 to -3.07 and a kurtosis ranging from 7.47 to 9.91.

kurtosis of data (Hu, Bentler, & Kano, 1992). Researchers have shown that the SB chi-square is a reliable test statistic for MACS analysis under various distributions and sample sizes (Curran, et al., 1995; Hu, et al., 1992). Satorra and Bentler (2001) further demonstrated how to calculate SB chi-square differences and corresponding degrees of freedom that are suitable for nested-model comparisons. An alternative to scaling the test statistic is to use a robust estimation method such as weighted least square (WLS) and robust WLS (RWLS).[12] These methods use the polychoric correlations, item means, and weight matrix to produce an asymptotic covariance matrix, which in turn is used to estimate the loading and intercept parameters (Muthén & Satorra, 1995).

Second, the overall model fit was not acceptable in some cases; the CFI values were quite low in conditions involving the 12-item scale. Note that the CFI depends on the average size of the correlations among observed variables (Bollen & Long, 1993). That is, if the average correlation is not high, then the CFI value will not be very high. In this study, the last six items of the 12-item scale had relatively small loadings by design, compared to the first six items of the same scale. Accordingly, the average correlation among the items decreased from the 6-item scale (0.57) to the 12-item scale (0.39). Thus, it appears that the small loadings resulted in the low CFI

[12] In fact, WLS estimation is not recommended for relatively small sample sizes. Flora and Curran (2004) noted that the chi-square is inflated, as are the parameter estimates, whereas their standard errors are negatively biased. Additionally, French and Finch (2006) found that the LR test with RWLS estimation provides very low power for testing metric invariance of a scale.

values for the 12-item scale.[13] In fact, these small loadings are somewhat smaller than those used in previous simulation studies (e.g., 0.60 in French & Finch, 2006, 2008; 0.48 - 0.66 in González-Romá et al., 2006; 0.50 - 0.80 in Kaplan, 1989; 0.58 - 0.90 in Stark et al., 2006).

Third, given that this is a Monte Carlo study, caution should be used in generalizing the results and conclusions beyond the conditions investigated. For example, the current study assumed no missing values in the item response. The conclusions of any DIF analysis likely depend on the amounts and the patterns of missing values. In addition, sample sizes were selected in this study so as to represent those often seen in the educational and psychological assessment. In some cases, however, smaller samples (i.e., less than 100) may be encountered, especially with low-incidence populations. Finally, the scales were relatively short, having 6 or 12 items, and only one or two items included DIF. Previous simulation studies found that the MACS technique performs better with larger scales and with smaller proportions of biased item in a scale (e.g., Finch, 2005; Meade & Lautenschlager, 2004; Navas-Ara & Gómez-Benito, 2002; Stark et al., 2006).

Taken together, further simulation work is encouraged to continue to examine the MACS technique under various additional conditions, as there are several problems that remain to be resolved in practice. These conditions may include

[13] The covariance/correlation between two measured variables can be obtained by $\lambda_{i1}\Psi\lambda_{i2}$ in the common factor model. Thus, the magnitude of the covariance/correlation depends on the magnitude of the loadings at a given trait variance.

normality/non-normal of data, estimation method, missing data, small sample size, scale size, and degree of contamination.

The current study found that the proposed ΔCFI criterion values are not optimal for testing measurement invariance, at least at the item level (see also French & Finch, 2006). As compared to IRT, one of the advantageous features of using CFA is to provide a variety of practical fit measures. Thus, future efforts are needed to empirically examine various fit indices and then find the criterion values that are suitable for DIF analysis. This would increase the utility of the MACS technique and expand the area to which the CFA approach for DIF analysis is applied. Following Cheung and Rensvold (2002), new criterion values should be independent of the overall fit of the baseline model, should not be influenced by model complexity, and should not be redundant with other fit indices.

Although most IRT models are based on the unidimensionality assumption, educational and psychological assessment often involves multidimensional surveys. For example, a test such as a licensure exam may measure several subsets of a skill. Accordingly, Raju and colleagues have proposed procedural guidelins as well as test statistics that are useful for assessing DIF in scales developed with multidimensional IRT models (e.g., Oshima, et al., 1997; Raju, et al., 1992). To my knowledge, however, no comparable DIF analysis has been proposed in the CFA literature. This lack of CFA methodology is clearly an area for additional future research.

6. Novel Contribution and Conclusion

This dissertation contributes to the measurement literature by cautioning researchers against the use of the conventional scaling method in case of DIF analysis. Because extensive prior research needed to establish a designated anchor set is rarely, if ever, carried out in the real world (Woods, 2009; see Thissen et al., 1993), it is likely that a researcher would innocently choose the conventional, marker-variable scaling method without realizing that it may lead to inflated Type I error. Consequently, it would be difficult to determine which items function truly differently from those that are falsely identified as having DIF.

Based on a simulation study, this dissertation suggests that, if used with the fixed-factor scaling method, Bonferroni-corrected LR test, and comparable large groups (e.g., greater than 100), the MACS technique would be a nearly fail-safe methodology for testing (at least uniform) DIF, even when a designated anchor set is not readily available. If properly followed, the recommended invariance-testing procedure provides accurate latent trait estimates for each group, thus making meaningful group comparisons tenable. Of course, the choice of which strategy to use must remain the prerogative of researchers. Hopefully, they may find the current findings and procedural guidance to be helpful in gaining a better understanding of invariance testing.

References

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Hillsdale, NJ: Erlbaum.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M. (2006). *EQS 6 structural equations program manual.* Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika,37,* 29–51.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Bollen, K. A. & Barb, K. H. (1981). Pearson's r and coarsely categorized measures. *American Sociological Review, 46*, 232-239.

Bollen, K.A. & Long, J. S. (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

Brown, T. A. (2006). *Confirmatory Factor Analysis for applied research*. New York: The Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Lawrence Erlbaum Associates.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*, 287-321.

Camilli, G., & Shepard, L. A. (1994). *Measurement methods for the social sciences series: Methods for identifying biased test items* (Vol. 4). Thousand Oaks, CA: Sage.

Chan, D. (2000). Detection of Differential Item Functioning on the Kirton Adaptation-Innovation Inventory using multiple-group Mean and Covariance Structure analyses. *Multivariate Behavioral Research, 35*, 169–199.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method, *Journal of Management, 25*, 1-27.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equation modeling. *Journal of Cross-Cultural Psychology, 31,* 187–212.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255.

Cohen, A. S., Kim, S. H., & Baker, F. B. (1993). Detection of Differential Item Functioning in the graded response model. *Applied Psychological Measurement, 17*, 335-350.

Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of Differential Item Functioning. *Applied Psychological Measurement, 20,* 15-26.

Curran, P. J. (1994). The robustness of confirmatory factor analysis to model misspecification and violations of normality. *Dissertation Abstracts International 55, 3-B,* 1220.

Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research, 18,* 147–167.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47,* 309-326.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are central issues. *Psychological Bulletin, 95,* 135-135.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70,* 662-680.

Everson, H. T., Millsap, R. E., & Rodriguez, C. M. (1991). Isolating gender differences in test anxiety: A Confirmatory Factor Analysis of the test anxiety inventory. *Educational and Psychological Measurement, 51,* 243-251.

Ferrando, P. J. (1996). Calibration of invariant item parameters in a continuous item response model using the extended LISREL measurement submodel. *Multivariate Behavioral Research, 31,* 419-439.

Fleishman, J. A. (2005). Using MIMIC models to assess the influence of differential item functioning. Retrieved October, 24 2005, from http://outcomes.cancer.gov/conference/irt/fleishman.pdf

Fleishman, J. A. & Lawrence, W. F. (2003). Demographic variation in SF-12 scores: True differences or differential item functioning. *Medical Care, 41,* 75-86.

Fleishman, J. A., Spector, W. D., & Altman, B. M. (2002). Impact of differential item functioning on age and gender differences in functional disability. J*ournal of Gerontology: Social Sciences, 57,* 275-283.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST and the IRT likelihood ratio test. *Applied Psychological Measurement, 29*, 278 –295.

French, B. F., & Finch, H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling, 13*, 378-402.

French, B. F., & Finch, H. (2008). Multigroup confirmatory factor analysis: locating the invariant referent sets. *Structural Equation Modeling, 15*, 96-113.

González, R., & Griffin, D. (2001). Testing parameters in structural equation modeling: Every "one" matters. *Psychological Methods, 6*, 258-269.

González-Romá, V., Hernández, A., & Gómez-Benito, J. (2006). Power and Type I error of the Mean and Covariance Structure analysis model for detecting Differential Item Functioning in graded response items. *Multivariate Behavioral Research, 41*, 29-53.

González-Romá, V., Tomás, I., Ferreres, D., & Hernández, A. (2005). Do items that measure self-perceived physical appearance function differentially across gender groups of adolescents? An application of the MACS model. *Structural Equation Modeling, 12*, 157–171.

Grayson, D. A., Mackinnon, A., Jorm, A. F., Creasey, H., & Broe, G. A. (2000). Item bias in the Center for Epidemiological Studies Depression Scale: Effects of physical disorders and disability in an elderly community sample. *Journal of Gerontology: Psychological Sciences: 55B*, 273-282.

Grayson, D. A., & Marsh, H. W. (1994). Identification with deficient rank loading matrices in Confirmatory Factor Analysis: Multitrait-multimethod models. *Psychometrika, 59,* 121-134.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the Confirmatory Factor Analysis framework. *Medical Care, 44,* 78-94.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hambleton, R. K., & Van der Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied P.sychologica1 Measurement, 6,* 373-378.

Helmstadter, G. (1964). *Principles of psychological measurement.* New York: Appleton-Century-Crofts.

Hernández, A., & González-Romá, V. (2003). Evaluating the multiple-group Mean and Covariance Structure analysis model for the detection of Differential Item Functioning in polytomous ordered items. *Psichotema, 15,* 322–327.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117-144.

Horn, J. L., & McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's book at the ethereal concept of factor invariance. *The Southern Psychologist, 1,* 179-188.

Hu, L.T., Bentler, B. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351–362.

Jones, R. N. (2006). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: Detecting Differential Item Functioning using MIMIC modeling. *Medical Care, 44*, 124-133.

Jöreskog, K. G. (1971a). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409-426.

Jöreskog, K.G. (1971b). Statistical analysis of sets of congeneric tests, *Psychometrika, 36*, 109-133.

Jöreskog, K.G. (1973). Analysis of covariance structures. In P. R. Krishnaiah (Ed.), *Multivariate analysis-III* (pp. 263-138). New York, NY: Academic Press.

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 10*, 631-639.

Jöreskog, K. G. & Sörbom, D. (1989). *LISREL 7*: *A guide to the program and applications*. Chicago: SPSS Publications.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Chicago, IL: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: Scientific Software International, Inc.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15*, 136-153.

Kaplan, D. (1989). Power of the likelihood ratio test in multiple group confirmatory factor-analysis under partial measurement invariance. *Educational and Psychological Measurement, 49*, 579-586.

Kaplan, D. & George, R. (1995). A study of power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling, 2*, 101-118.

Kim, S. H., Cohen, A. S., & Park, T. H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*, 261–276

Lawley, D. N. (1943-44). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh, 62*, 28-30.

Little, T. D. (1997). Mean and Covariance Structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32,* 53-76.

Little, T. D. (in press). *Longitudinal SEM.* New York, NY: Guilford Press.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling, 13*, 59-72.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful

group comparisons. *Structural Equation Modeling, 11*, 514-534.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillside, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison Wesley.

Luijben, T. C., Boomsma, A., & Molenaar, I. W. (1988). Modification of a factor analysis model in covariance structure analysis: A Monte Carlo study. In T. K. Dijkstra (Ed.), *On model uncertainty and its statistical implications*. Berlin: Springer-Verlag.

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100*, 107-120.

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common factor model: Implications for theory and practice. *Psychological Bulletin, 109*, 502-511.

Marsh, H. W. (1994). Confirmatory Factor Analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling, 1*, 5-34.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness of fit indexes in Confirmatory Factor Analysis: The effect of sample size. *Psychological Bulletin, 103*, 391-410.

Marsh, H. W., Hey, J., & Roche, L. A. (1997). Structure of physical self-concept: Elite athletes and physical education students. *Journal of Educational Psychology, 89*, 369–380.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika,47,* 149–174.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research, 41,* 55–64.

McArdle, J. J., & McDonald, R. P. (1984). Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology, 37,* 234-251.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification, 6,* 97–103.

McDonald, R. P. (1999). *Test theory. Unified treatment.* Mahwah, NJ: Lawrence Erlbaum.

Meade, A. W., & Lautenschlager, G. K. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling, 11,* 60–72.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29,* 223-237.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525-543.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44,* 69-77.

Muraki, E. (1990). Fitting polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14,* 59–71.

Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainder & H. Braun (Eds.), *Test validity* (pp. 213-238). Hillsdale, NJ: Lawrence Erlbaum.

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus*. Los Angeles: University of California and Muthén & Muthén.

Muthén, B. O., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46*, 407-419.

Muthén, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*, 1-22.

Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus user's guide*. Fifth edition. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., & Lehman, J. (1988). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10*, 133-142.

Navas-Ara, M. J., & Gomez-Benito, J. (2002). Effects of ability scale purification on identification of DIF. *European Journal of Psychological Assessment, 18*, 9-15.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5*, 107–124.

Oshima, T. C., Raju, N. S., & Flowers, C. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functioning of items and tests. *Journal of Educational Measurement,34*, 253–272.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrik*a, 53, 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14*, 197-207.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on Confirmatory Factor Analysis and item response theory. *Journal of Applied Psychology, 87,* 517–529.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory Factor Analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*, 552-566.

Raju, N. S., Van der Linden, W. J., & Fleer, P. F. (1992). *An IRT-based internal measure of test bias with applications for Differential Item Functioning.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Rock, D. A., Werts, C. E., & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research, 13*, 403-418.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs*, No. 17.

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monographs*, No. 18.

SAS Institute (2004). *SAS/STAT 9.1 user's guide*. Cary, NC: SAS Institute Inc.

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. *American Statistical Association 1988 proceedings of the business and economics section* (pp. 308–313). Alexandria VA: American Statistical Association.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507-514.

Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the ability distribution. *Applied Psychological Measurement, 14*, 299-311.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229-239.

Sörbom, D. (1982). Structural equation models with structured means. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation* (pp. 183-195). Amsterdam: North Holland.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1202-1306.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in crossnational consumer research. *Journal of Consumer Research, 25*, 78-90.

Steiger, J. H. (1989). *EzPATH: Causal modeling.* Evanston, IL: SYSTAT.

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika, 50*, 253-264.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1-16.

Takane, Y. & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

Teresi, J. A. (2006). Overview of quantitative measurement methods: Equivalence, invariance, and Differential Item Functioning in health applications. *Medical Care, 44*, 39-49.

Thissen, D. (2001). *IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for Differential Item*

Functioning. Chapel Hill: L. L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill. Scientific Software, Inc.; 1991.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99,*118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of Differential Item Functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods,* 5, 139-158.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–69.

Wang, W. C. (2004). Effects of anchor item methods on detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education, 72*, 221-261.

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*, 479-498.

Wanichtanom, R. (2001). *Methods of detecting Differential Item Functioning: A comparison of item response theory and Confirmatory Factor Analysis*. Unpublished doctoral dissertation

Wasti, S. A., Bergman, M. E., Glomb, T. M., & Drasgow, F. (2000). Test of the cross-cultural generalizability of a model of sexual harassment. *Journal of Applied Psychology, 85*, 766–778.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). Washington, DC: American Psychological Association.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning, *Applied Psychological Measurement, 33*, 42-57.

Yuan, K. H., & Bentler, P. M. (2004). On chi-square difference and *z* tests in Mean and Covariance Structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64*, 737–757.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Appendix A

Simulation Design

The present study investigates Type I error and power of the free-baseline MACS technique for detecting DIF. It has a 2 × 2 × 3 × 2 × 3 × 3 × 2 × 4 × 3 factorial design (5,184 cells): two types of item response, by two scale sizes, by three combinations of sample sizes, by two combinations of latent trait distributions, by three types of anchor item, by three types of target item, by two amounts of DIF in the target item (if present), by four criterion values for testing DIF, and by three scaling methods.

Item response	Scale size	Sample size	Trait Distribution	DIF in anchor	DIF in target	DIF size	Criterion value	Scaling method
Dichotomous	6 items	100 : 900	$N[0,1]$	No	No	Small	Uncorrected *p*	Marker-variable
Polytomous	12 items	250 : 750	$N[-1,1]$	Uniform	Uniform	Large	Corrected *p*	Fixed-factor
		500 : 500		Non-uniform	Non-uniform		–0.01 ΔCFI	Effects-coded
							–0.02 ΔCFI	

Appendix B

Type I Error

The uncorrected *p* values of the LR test were .05 in all conditions. The corrected *p* values were .01 and .005 in the 6-item and 12-item conditions, respectively.

Type I Error of Using Marker-Variable Scaling Method for Detecting Non-Uniform DIF in the Equal Latent Trait Mean Condition

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Dichotomous	6 items	100:900	DIF-free	.012	.007	0	0
			Non-uniform	.890	.873	.659	.880
			Uniform	1	.993	0	.586
		250:750	DIF-free	.012	.002	0	.014
			Non-uniform	.327	.192	0	.045
			Uniform	1	1	.006	.998
		500:500	DIF-free	.010	0	0	0
			Non-uniform	.286	.154	0	.017
			Uniform	1	1	.064	1
	12 items	100:900	DIF-free	.040	.013	0	0
			Non-uniform	.886	.840	.138	.188
			Uniform	.988	.972	.614	.795
		250:750	DIF-free	.046	.014	0	0
			Non-uniform	.913	.874	0	.606
			Uniform	.998	.986	0	.557
		500:500	DIF-free	.050	.002	0	0
			Non-uniform	.952	.926	0	.868
			Uniform	1	.998	0	.894

Type I Error of Using Marker-Variable Scaling Method for Detecting Non-Uniform DIF in the Equal Latent Trait Mean Condition (Continued)

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	.008	0	0	0
			Non-uniform	.141	.048	0	0
			Uniform	.042	.012	0	0
		250:750	DIF-free	0	0	0	0
			Non-uniform	.134	.038	0	0
			Uniform	.068	.010	0	0
		500:500	DIF-free	.002	0	0	0
			Non-uniform	.096	.032	0	0
			Uniform	.114	.020	0	0
	12 items	100:900	DIF-free	.022	.004	0	0
			Non-uniform	.852	.782	.217	.366
			Uniform	.062	.006	0	0
		250:750	DIF-free	.002	0	0	0
			Non-uniform	.838	.798	0	.002
			Uniform	.132	.004	0	0
		500:500	DIF-free	.002	0	0	0
			Non-uniform	.932	.920	0	.048
			Uniform	.178	.020	0	0

Type I Error of Using Marker-Variable Scaling Method for Detecting Uniform DIF in the Equal Latent Trait Mean Condition

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Dichotomous	6 items	100:900	DIF-free	.005	.005	0	0
			Non-uniform	.062	.010	0	0
			Uniform	1	.988	.612	.927
		250:750	DIF-free	.016	0	0	0
			Non-uniform	.078	.012	0	0
			Uniform	1	1	.690	1
		500:500	DIF-free	.024	.002	0	0
			Non-uniform	.042	.012	0	0
			Uniform	1	1	1	1
	12 items	100:900	DIF-free	.003	0	0	0
			Non-uniform	.656	.630	.551	.760
			Uniform	1	1	.315	.619
		250:750	DIF-free	.028	0	0	0
			Non-uniform	.093	.012	0	0
			Uniform	1	1	0	1
		500:500	DIF-free	.028	0	0	0
			Non-uniform	.056	.010	0	0
			Uniform	1	1	.002	1

Type I Error of Using Marker-Variable Scaling Method for Detecting Uniform DIF in the Equal Latent Trait Mean Condition (Continued)

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	.004	.002	0	0
			Non-uniform	.051	.016	0	0
			Uniform	1	1	.165	1
		250:750	DIF-free	0	0	0	0
			Non-uniform	.080	.020	0	0
			Uniform	1	1	1	1
		500:500	DIF-free	0	0	0	0
			Non-uniform	.040	.008	0	0
			Uniform	1	1	1	1
	12 items	100:900	DIF-free	.004	0	0	0
			Non-uniform	.066	.012	0	.006
			Uniform	.998	.998	0	.988
		250:750	DIF-free	.002	0	0	0
			Non-uniform	.080	.010	0	0
			Uniform	1	1	.064	1
		500:500	DIF-free	0	0	0	0
			Non-uniform	.046	.008	0	0
			Uniform	1	1	.998	1

Type I Error of Using Fixed-Factor Scaling Method for Detecting Non-Uniform DIF in the Equal Latent Trait Mean Condition

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Dichotomous	6 items	100:900	DIF-free	.023	.021	0	.060
			Non-uniform	.197	.186	.378	.459
			Uniform	.005	0	0	0
		250:750	DIF-free	.022	.004	0	0
			Non-uniform	.062	.008	.006	.012
			Uniform	.028	0	0	0
		500:500	DIF-free	.020	.008	0	0
			Non-uniform	.056	.008	0	0
			Uniform	.026	.004	0	0
	12 items	100:900	DIF-free	.055	.025	0	0
			Non-uniform	.986	.968	.716	.898
			Uniform	.465	.420	.452	.606
		250:750	DIF-free	.044	.016	0	.007
			Non-uniform	.760	.612	0	.002
			Uniform	.098	.052	0	.023
		500:500	DIF-free	.046	.014	0	0
			Non-uniform	.836	.778	0	.007
			Uniform	.078	.024	0	.011

Type I Error of Using Fixed-Factor Scaling Method for Detecting Non-Uniform DIF in the Equal Latent Trait Mean Condition (Continued)

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	.020	.004	0	0
			Non-uniform	.010	.002	0	0
			Uniform	.014	.002	0	0
		250:750	DIF-free	.008	0	0	0
			Non-uniform	.002	0	0	0
			Uniform	.004	0	0	0
		500:500	DIF-free	.040	.008	0	0
			Non-uniform	.018	0	0	0
			Uniform	.032	.004	0	0
	12 items	100:900	DIF-free	.024	.002	0	0
			Non-uniform	.395	.270	.182	.327
			Uniform	.022	.002	0	0
		250:750	DIF-free	.012	0	0	0
			Non-uniform	.462	.162	0	.005
			Uniform	.002	0	0	0
		500:500	DIF-free	.046	0	0	0
			Non-uniform	.591	.232	0	0
			Uniform	.030	0	0	0

Type I Error of Using Fixed-Factor Scaling Method for Detecting Uniform DIF in the Equal Latent Trait Mean Condition

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Dichotomous	6 items	100:900	DIF-free	0	0	0	0
			Non-uniform	.372	.372	.560	.828
			Uniform	.002	.002	0	.019
		250:750	DIF-free	.166	.016	0	0
			Non-uniform	.162	.014	0	0
			Uniform	.166	.014	0	0
		500:500	DIF-free	.136	.004	0	0
			Non-uniform	.132	.004	0	0
			Uniform	.132	.004	0	0
	12 items	100:900	DIF-free	.005	.005	0	.019
			Non-uniform	.574	.574	.672	.867
			Uniform	.413	.413	.611	.802
		250:750	DIF-free	.162	.006	0	0
			Non-uniform	.110	0	0	0
			Uniform	.156	.006	0	0
		500:500	DIF-free	.136	.002	0	0
			Non-uniform	.114	.002	0	0
			Uniform	.132	.002	0	0

Type I Error of Using Fixed-Factor Scaling Method for Detecting Uniform DIF in the Equal Latent Trait Mean Condition (Continued)

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	0	0	0	0
			Non-uniform	.004	.004	0	0
			Uniform	0	0	0	0
		250:750	DIF-free	.024	.002	0	0
			Non-uniform	.024	.002	0	0
			Uniform	.024	.002	0	0
		500:500	DIF-free	.006	0	0	0
			Non-uniform	.006	0	0	0
			Uniform	.006	0	0	0
	12 items	100:900	DIF-free	0	0	0	0
			Non-uniform	.405	.394	.503	.848
			Uniform	.002	.002	0	0
		250:750	DIF-free	.024	0	0	0
			Non-uniform	.016	.006	0	.013
			Uniform	.026	.002	0	.004
		500:500	DIF-free	.006	0	0	0
			Non-uniform	.008	.002	0	0
			Uniform	.006	0	0	0

Type I Error of Using Effects-Coded Scaling Method for Detecting Non-Uniform DIF in the Equal Latent Trait Mean Condition

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Dichotomous	6 items	100:900	DIF-free	.830	.799	.660	.868
			Non-uniform	.909	.874	.677	.836
			Uniform	.981	.975	.701	.884
		250:750	DIF-free	.234	.122	0	.029
			Non-uniform	.612	.494	.314	.463
			Uniform	.914	.895	.626	.872
		500:500	DIF-free	.284	.137	0	.007
			Non-uniform	.360	.232	0	.032
			Uniform	.465	.313	0	.073
	12 items	100:900	DIF-free	.944	.934	.642	.886
			Non-uniform	.941	.905	.522	.654
			Uniform	1	.995	.701	.943
		250:750	DIF-free	.411	.232	0	.023
			Non-uniform	.902	.863	.011	.054
			Uniform	.728	.647	.451	.620
		500:500	DIF-free	.428	.254	0	.006
			Non-uniform	.951	.922	.009	.388
			Uniform	.384	.254	0	.012

Type I Error of Using Effects-Coded Scaling Method for Detecting Non-Uniform DIF in the Equal Latent Trait Mean Condition (Continued)

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	.115	.048	0	0
			Non-uniform	.135	.044	0	0
			Uniform	.105	.040	0	.005
		250:750	DIF-free	.092	.026	0	0
			Non-uniform	.164	.050	0	0
			Uniform	.086	.028	0	0
		500:500	DIF-free	.100	.034	0	0
			Non-uniform	.160	.058	0	0
			Uniform	.084	.030	0	0
	12 items	100:900	DIF-free	.223	.080	0	.004
			Non-uniform	.595	.284	0	.012
			Uniform	.177	.048	0	0
		250:750	DIF-free	.184	.056	0	0
			Non-uniform	.856	.719	0	0
			Uniform	.148	.036	0	0
		500:500	DIF-free	.178	.054	0	0
			Non-uniform	.901	.847	0	.009
			Uniform	.148	.026	0	0

Type I Error of Using Effects-Coded Scaling Method for Detecting Uniform DIF in the Equal Latent Trait Mean Condition

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Dichotomous	6 items	100:900	DIF-free	.661	.654	.618	.827
			Non-uniform	.732	.711	.601	.828
			Uniform	.958	.940	.645	.892
		250:750	DIF-free	.050	.008	0	0
			Non-uniform	.112	.018	0	0
			Uniform	.812	.614	.148	.251
		500:500	DIF-free	.032	.008	0	.007
			Non-uniform	.068	.010	0	0
			Uniform	.968	.850	0	.198
	12 items	100:900	DIF-free	.986	.981	.626	.944
			Non-uniform	.889	.876	.706	.912
			Uniform	.867	.836	.670	.855
		250:750	DIF-free	.099	.018	0	.005
			Non-uniform	.081	.040	0	.052
			Uniform	.965	.958	.683	.929
		500:500	DIF-free	.060	.004	0	0
			Non-uniform	.071	.037	0	.082
			Uniform	.285	.084	0	0

Type I Error of Using Effects-Coded Scaling Method for Detecting Uniform DIF in the Equal Latent Trait Mean Condition (Continued)

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	.030	.004	0	0
			Non-uniform	.042	.010	0	0
			Uniform	.978	.898	0	.193
		250:750	DIF-free	.038	.010	0	0
			Non-uniform	.062	.018	0	0
			Uniform	1	.998	0	.708
		500:500	DIF-free	.040	.006	0	0
			Non-uniform	.050	.004	0	0
			Uniform	1	1	0	.962
	12 items	100:900	DIF-free	.064	.008	0	0
			Non-uniform	.031	.010	0	.018
			Uniform	.452	.171	0	0
		250:750	DIF-free	.104	.012	0	0
			Non-uniform	.133	.095	.171	.256
			Uniform	.576	.242	0	0
		500:500	DIF-free	.068	.008	0	0
			Non-uniform	.032	.006	0	.006
			Uniform	.762	.392	0	0

Type I Error of Using Marker-Variable Scaling Method for Detecting Non-Uniform DIF in the Unequal Latent Trait Mean Condition

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Dichotomous	6 items	100:900	DIF-free	.102	.036	0	.013
			Non-uniform	.575	.509	.230	.396
			Uniform	.991	.974	.714	.936
		250:750	DIF-free	.090	.031	0	0
			Non-uniform	.611	.442	0	.149
			Uniform	1	1	.843	1
		500:500	DIF-free	.102	.038	0	0
			Non-uniform	.782	.640	0	.156
			Uniform	1	1	1	1
	12 items	100:900	DIF-free	.617	.565	.518	.645
			Non-uniform	.984	.963	.577	.869
			Uniform	1	.995	.726	.949
		250:750	DIF-free	.170	.076	0	0
			Non-uniform	.898	.861	.236	.716
			Uniform	.990	.990	.376	.970
		500:500	DIF-free	.130	.044	0	0
			Non-uniform	.922	.873	0	.660
			Uniform	1	1	.569	.996

Type I Error of Using Marker-Variable Scaling Method for Detecting Non-Uniform DIF in the Unequal Latent Trait Mean Condition (Continued)

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	0	0	0	0
			Non-uniform	.113	.038	0	0
			Uniform	.795	.475	0	.006
		250:750	DIF-free	0	0	0	0
			Non-uniform	.092	.036	0	.005
			Uniform	.958	.820	0	.020
		500:500	DIF-free	0	0	0	0
			Non-uniform	.098	.032	0	0
			Uniform	1	.980	0	.172
	12 items	100:900	DIF-free	.004	0	0	0
			Non-uniform	.746	.586	0	.002
			Uniform	.821	.400	0	0
		250:750	DIF-free	0	0	0	0
			Non-uniform	.821	.745	0	0
			Uniform	.982	.792	0	0
		500:500	DIF-free	0	0	0	0
			Non-uniform	.923	.899	0	.059
			Uniform	1	.974	0	0

Type I Error of Using Marker-Variable Scaling Method for Detecting Uniform DIF in the Unequal Latent Trait Mean Condition

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Dichotomous	6 items	100:900	DIF-free	.946	.914	.598	.854
			Non-uniform	.489	.208	0	.006
			Uniform	.997	.990	.382	.945
		250:750	DIF-free	.914	.713	0	.059
			Non-uniform	.692	.446	0	.032
			Uniform	1	1	.339	1
		500:500	DIF-free	.990	.918	0	.154
			Non-uniform	.910	.733	0	.103
			Uniform	1	1	1	1
	12 items	100:900	DIF-free	.639	.248	0	0
			Non-uniform	.621	.218	0	.006
			Uniform	.992	.988	.631	.752
		250:750	DIF-free	.867	.559	0	.002
			Non-uniform	.875	.632	0	0
			Uniform	.998	.998	.128	.974
		500:500	DIF-free	.966	.778	0	0
			Non-uniform	.972	.874	0	.142
			Uniform	1	1	.458	1

Type I Error of Using Marker-Variable Scaling Method for Detecting Uniform DIF in the Unequal Latent Trait Mean Condition (Continued)

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	.608	.229	0	0
			Non-uniform	.940	.833	0	.132
			Uniform	1	1	.606	1
		250:750	DIF-free	.956	.762	0	.004
			Non-uniform	1	1	0	.792
			Uniform	1	1	1	1
		500:500	DIF-free	.992	.918	0	.034
			Non-uniform	1	1	0	.926
			Uniform	1	1	1	1
	12 items	100:900	DIF-free	.602	.141	0	0
			Non-uniform	.964	.886	.324	.485
			Uniform	1	1	0	1
		250:750	DIF-free	.942	.596	0	0
			Non-uniform	1	1	0	.228
			Uniform	1	1	.314	1
		500:500	DIF-free	.986	.824	0	0
			Non-uniform	1	1	0	.808
			Uniform	1	1	1	1

Type I Error of Using Fixed-Factor Scaling Method for Detecting Non-Uniform DIF in the Unequal Latent Trait Mean Condition

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Dichotomous	6 items	100:900	DIF-free	.944	.913	.542	.796
			Non-uniform	.935	.848	.376	.571
			Uniform	.970	.923	.525	.756
		250:750	DIF-free	.998	.984	0	.559
			Non-uniform	1	.994	0	.695
			Uniform	1	.988	.004	.668
		500:500	DIF-free	1	.994	0	.752
			Non-uniform	1	1	0	.876
			Uniform	1	1	0	.888
	12 items	100:900	DIF-free	.885	.716	.282	.369
			Non-uniform	.969	.912	.374	.531
			Uniform	.966	.919	.489	.692
		250:750	DIF-free	1	.963	0	.006
			Non-uniform	1	.984	0	.301
			Uniform	1	.974	0	.020
		500:500	DIF-free	.988	.954	0	.064
			Non-uniform	.996	.992	0	.469
			Uniform	.990	.970	0	.139

Type I Error of Using Fixed-Factor Scaling Method for Detecting Non-Uniform DIF in the Unequal Latent Trait Mean Condition (Continued)

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	0	0	0	0
			Non-uniform	.004	0	0	0
			Uniform	.002	.002	0	.015
		250:750	DIF-free	.006	0	0	0
			Non-uniform	.014	.002	0	0
			Uniform	.004	0	0	0
		500:500	DIF-free	.010	0	0	0
			Non-uniform	.012	0	0	0
			Uniform	.006	0	0	0
	12 items	100:900	DIF-free	.002	0	0	0
			Non-uniform	.203	.033	.010	.013
			Uniform	.002	0	0	0
		250:750	DIF-free	.008	0	0	0
			Non-uniform	.478	.172	0	.005
			Uniform	.004	0	0	0
		500:500	DIF-free	.008	0	0	0
			Non-uniform	.643	.302	0	0
			Uniform	.002	0	0	0

Type I Error of Using Fixed-Factor Scaling Method for Detecting Uniform DIF in the Unequal Latent Trait Mean Condition

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Dichotomous	6 items	100:900	DIF-free	1	1	.002	1
			Non-uniform	.986	.973	.429	.965
			Uniform	.988	.983	.226	.991
		250:750	DIF-free	1	1	.644	1
			Non-uniform	1	1	.516	1
			Uniform	1	1	.679	1
		500:500	DIF-free	1	1	1	1
			Non-uniform	1	1	.998	1
			Uniform	1	1	1	1
	12 items	100:900	DIF-free	1	1	.308	.471
			Non-uniform	.996	.996	.562	.748
			Uniform	.993	.990	.292	.495
		250:750	DIF-free	1	1	0	1
			Non-uniform	.998	.996	0	.992
			Uniform	1	1	0	1
		500:500	DIF-free	1	1	0	1
			Non-uniform	1	1	0	.998
			Uniform	1	1	0	1

Type I Error of Using Fixed-Factor Scaling Method for Detecting Uniform DIF in the Unequal Latent Trait Mean Condition (Continued)

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	1	1	0	1
			Non-uniform	1	1	0	1
			Uniform	1	1	0	1
		250:750	DIF-free	1	1	1	1
			Non-uniform	1	1	.998	1
			Uniform	1	1	1	1
		500:500	DIF-free	1	1	1	1
			Non-uniform	1	1	1	1
			Uniform	1	1	1	1
	12 items	100:900	DIF-free	1	1	0	.751
			Non-uniform	1	1	0	.368
			Uniform	1	1	0	.728
		250:750	DIF-free	1	1	0	1
			Non-uniform	1	1	0	1
			Uniform	1	1	0	1
		500:500	DIF-free	1	1	0	1
			Non-uniform	1	1	0	1
			Uniform	1	1	0	1

Type I Error of Using Effects-Coded Scaling Method for Detecting Non-Uniform DIF in the Unequal Latent Trait Mean Condition

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Dichotomous	6 items	100:900	DIF-free	.921	.882	.665	.863
			Non-uniform	.987	.981	.745	.968
			Uniform	.966	.959	.612	.935
		250:750	DIF-free	.400	.251	.009	.031
			Non-uniform	.881	.814	.495	.686
			Uniform	.980	.955	.615	.902
		500:500	DIF-free	.368	.252	.006	.080
			Non-uniform	.792	.660	0	.324
			Uniform	.464	.347	.057	.216
	12 items	100:900	DIF-free	1	.986	.744	.976
			Non-uniform	1	.995	.682	.919
			Uniform	.987	.987	.683	.910
		250:750	DIF-free	.985	.981	.689	.942
			Non-uniform	.850	.777	.002	.446
			Uniform	.889	.854	.573	.816
		500:500	DIF-free	.474	.290	0	.039
			Non-uniform	.908	.856	0	.635
			Uniform	.485	.347	.085	.146

Type I Error of Using Effects-Coded Scaling Method for Detecting Non-Uniform DIF in the Unequal Latent Trait Mean Condition (Continued)

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	.103	.030	0	.010
			Non-uniform	.155	.063	0	.004
			Uniform	.175	.082	0	.009
		250:750	DIF-free	.094	.040	0	.005
			Non-uniform	.154	.048	0	.004
			Uniform	.150	.074	0	.013
		500:500	DIF-free	.064	.018	0	0
			Non-uniform	.204	.074	0	.004
			Uniform	.196	.078	0	.008
	12 items	100:900	DIF-free	.203	.105	0	.004
			Non-uniform	.631	.314	0	.010
			Uniform	.183	.074	0	.004
		250:750	DIF-free	.168	.056	0	0
			Non-uniform	.836	.670	0	0
			Uniform	.166	.040	0	.004
		500:500	DIF-free	.142	.038	0	0
			Non-uniform	.900	.854	0	.008
			Uniform	.142	.028	0	0

Type I Error of Using Effects-Coded Scaling Method for Detecting Uniform DIF in the Unequal Latent Trait Mean Condition

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Dichotomous	6 items	100:900	DIF-free	.859	.791	.616	.814
			Non-uniform	.981	.981	.725	.931
			Uniform	.887	.871	.615	.869
		250:750	DIF-free	.441	.247	.005	.019
			Non-uniform	.935	.806	.002	.210
			Uniform	.969	.952	.672	.926
		500:500	DIF-free	.571	.371	0	.032
			Non-uniform	.992	.976	0	.558
			Uniform	.801	.777	.668	.865
	12 items	100:900	DIF-free	.768	.633	.366	.480
			Non-uniform	.797	.688	.456	.573
			Uniform	.932	.889	.649	.891
		250:750	DIF-free	.996	.977	.693	.922
			Non-uniform	.900	.724	.214	.284
			Uniform	.703	.590	.405	.527
		500:500	DIF-free	.856	.688	.223	.274
			Non-uniform	.996	.962	.006	.036
			Uniform	.656	.441	.248	.316

Type I Error of Using Effects-Coded Scaling Method for Detecting Uniform DIF in the Unequal Latent Trait Mean Condition (Continued)

				Criterion value			
				LR test		ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*	Corrected *p*	–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	.475	.233	0	.005
			Non-uniform	.934	.819	0	.084
			Uniform	.370	.205	0	.014
		250:750	DIF-free	.866	.654	0	.035
			Non-uniform	1	1	0	.802
			Uniform	.550	.308	0	.009
		500:500	DIF-free	.956	.842	0	.080
			Non-uniform	1	1	0	.992
			Uniform	.702	.470	0	.032
	12 items	100:900	DIF-free	.925	.754	0	0
			Non-uniform	.915	.682	0	.008
			Uniform	.423	.139	0	0
		250:750	DIF-free	1	.992	0	.036
			Non-uniform	1	.998	0	.028
			Uniform	.836	.520	0	0
		500:500	DIF-free	1	1	0	.200
			Non-uniform	.998	.998	0	.466
			Uniform	.926	.738	0	0

Appendix C

Power

Power was not reported in the cases where Type I error was inflated at the nominal alpha level (.05). The values in parentheses indicate that they are for the “small DIF” conditions; adjacent values are for the “large DIF” conditions.

Power of Using Marker-Variable Scaling Method for Detecting Non-Uniform DIF in the Equal Latent Trait Mean Condition

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Dichotomous	6 items	100:900	DIF-free	(.349)	.294	(.133)	.146	(0)	0	(.008)	.012
			Non-uniform	(.892)	.967	(.882)	.834	(.670)	.268	(.895)	.451
			Uniform	(.996)	.813	(.982)	.591	(.699)	0	(.940)	.067
		250:750	DIF-free	(.552)	.521	(.356)	.335	(.003)	0	(.028)	.026
			Non-uniform	(.313)	1	(.182)	.994	(0)	.004	(.030)	.560
			Uniform	(1)	.978	(1)	.882	(.044)	0	(1)	.384
		500:500	DIF-free	(.640)	.744	(.412)	.564	(0)	0	(.022)	.120
			Non-uniform	(.308)	1	(.116)	1	(0)	.008	(0)	.753
			Uniform	(1)	.996	(1)	.986	(.364)	0	(1)	.688
	12 items	100:900	DIF-free	(.710)	.804	(.465)	.724	(.257)	.353	(.317)	.464
			Non-uniform	(.910)	.914	(.881)	.843	(.574)	.500	(.749)	.626
			Uniform	(1)	.991	(1)	.968	(.572)	.578	(.763)	.840
		250:750	DIF-free	(.786)	.854	(.558)	.784	(0)	.076	(.002)	.256
			Non-uniform	(.849)	.789	(.682)	.625	(0)	.144	(0)	.206
			Uniform	(1)	.996	(1)	.974	(0)	.043	(.984)	.803
		500:500	DIF-free	(.836)	.902	(.616)	.841	(0)	.004	(.010)	.581
			Non-uniform	(.900)	.802	(.722)	.500	(0)	0	(0)	0
			Uniform	(1)	1	(1)	1	(0)	.040	(.990)	.924

Power of Using Marker-Variable Scaling Method for Detecting Non-Uniform DIF in the Equal Latent Trait Mean Condition

(Continued)

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Polytomous	6 items	100:900	DIF-free	(.183)	.167	(.051)	.048	(0)	0	(.004)	0
			Non-uniform	(.081)	.264	(.024)	.108	(0)	0	(.006)	.016
			Uniform	(.323)	.378	(.154)	.342	(0)	.363	(.003)	.552
		250:750	DIF-free	(.289)	.272	(.096)	.112	(0)	0	(0)	0
			Non-uniform	(.066)	.580	(.018)	.256	(0)	.006	(0)	.012
			Uniform	(.612)	.060	(.336)	.020	(0)	0	(0)	0
		500:500	DIF-free	(.293)	.434	(.076)	.198	(0)	0	(0)	0
			Non-uniform	(.050)	.778	(.004)	.396	(0)	0	(0)	.004
			Uniform	(.780)	.060	(.500)	.008	(0)	0	(0)	0
	12 items	100:900	DIF-free	(.326)	.963	(.079)	.952	(0)	.446	(.003)	.815
			Non-uniform	(.582)	.182	(.384)	.042	(.155)	0	(.270)	0
			Uniform	(.557)	.963	(.192)	.945	(0)	.471	(.002)	.813
		250:750	DIF-free	(.578)	.838	(.234)	.792	(0)	.002	(0)	.044
			Non-uniform	(.556)	.152	(.248)	.028	(0)	0	(0)	0
			Uniform	(.862)	.862	(.524)	.852	(0)	.007	(0)	.102
		500:500	DIF-free	(.664)	.930	(.266)	.920	(0)	.004	(0)	.278
			Non-uniform	(.654)	.090	(.286)	.004	(0)	0	(0)	0
			Uniform	(.968)	.942	(.784)	.938	(0)	.013	(0)	.675

Power of Using Marker-Variable Scaling Method for Detecting Uniform DIF in the Equal Latent Trait Mean Condition

				Criterion value					
				LR test				ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01	–0.002
Dichotomous	6 items	100:900	DIF-free	(.945)	1	(.732)	1	(.114) 0	(.153) .966
			Non-uniform	(.834)	1	(.619)	1	(0) 0	(.033) .913
			Uniform	(.997)	.199	(.997)	.043	(.117) 0	(.860) 0
		250:750	DIF-free	(.972)	1	(.840)	1	(0) 0	(.016) 1
			Non-uniform	(.996)	1	(.972)	1	(0) .004	(.386) 1
			Uniform	(1)	.323	(1)	.110	(0) .002	(1) .002
		500:500	DIF-free	(.998)	1	(.970)	1	(0) .064	(.154) 1
			Non-uniform	(1)	1	(.992)	1	(0) .254	(.592) 1
			Uniform	(1)	.312	(1)	.100	(.334) 0	(1) 0
	12 items	100:900	DIF-free	(.953)	1	(.604)	.995	(.140) .511	(.183) .660
			Non-uniform	(.852)	1	(.526)	1	(0) .325	(.003) .471
			Uniform	(.996)	.557	(.992)	.460	(.618) .376	(.862) .494
		250:750	DIF-free	(.974)	1	(.771)	1	(0) 0	(0) .633
			Non-uniform	(.998)	1	(.948)	1	(0) .002	(.010) .912
			Uniform	(1)	.280	(1)	.045	(0) 0	(.688) 0
		500:500	DIF-free	(1)	1	(.948)	1	(0) 0	(0) .984
			Non-uniform	(1)	1	(.994)	1	(0) 0	(.034) .996
			Uniform	(1)	.212	(1)	.026	(0) 0	(.998) 0

Power of Using Marker-Variable Scaling Method for Detecting Uniform DIF in the Equal Latent Trait Mean Condition

(Continued)

				Criterion value					
				LR test				ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Polytomous	6 items	100:900	DIF-free	(1)	1	(.992)	1	(0)	.038	(.178)	1
			Non-uniform	(.982)	1	(.926)	1	(0)	.069	(.182)	1
			Uniform	(1)	.008	(1)	0	(0)	0	(.998)	0
		250:750	DIF-free	(1)	1	(1)	1	(0)	1	(.944)	1
			Non-uniform	(1)	1	(1)	1	(0)	1	(.962)	1
			Uniform	(1)	.094	(1)	.010	(.404)	0	(1)	0
		500:500	DIF-free	(1)	1	(1)	1	(0)	1	(.998)	1
			Non-uniform	(1)	1	(1)	1	(0)	1	(.990)	1
			Uniform	(1)	.208	(1)	.028	(.998)	0	(1)	0
	12 items	100:900	DIF-free	(1)	.993	(.968)	.991	(0)	.138	(0)	.974
			Non-uniform	(.982)	.998	(.883)	.998	(0)	0	(0)	.976
			Uniform	(.997)	.012	(.994)	0	(.338)	0	(.614)	0
		250:750	DIF-free	(1)	1	(1)	1	(0)	0	(.012)	1
			Non-uniform	(1)	1	(1)	1	(0)	.046	(.242)	1
			Uniform	(1)	.084	(1)	.002	(0)	0	(1)	0
		500:500	DIF-free	(1)	1	(1)	1	(0)	.606	(.324)	1
			Non-uniform	(1)	1	(1)	1	(0)	.914	(.616)	1
			Uniform	(1)	.176	(1)	.006	(0)	0	(1)	0

Power of Using Fixed-Factor Scaling Method for Detecting Non-Uniform DIF in the Equal Latent Trait Mean Condition

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Dichotomous	6 items	100:900	DIF-free	(.726)	.430	(.704)	.150	(.628)	0	(.867)	0
			Non-uniform	(.839)	.495	(.829)	.436	(.684)	.387	(.898)	.542
			Uniform	(.216)	.360	(.204)	.144	(.363)	0	(.496)	.005
		250:750	DIF-free	(.024)	.886	(.006)	.692	(0)	0	(0)	.050
			Non-uniform	(.056)	.412	(.008)	.267	(0)	0	(.003)	.042
			Uniform	(.030)	.834	(.002)	.592	(0)	0	(0)	.039
		500:500	DIF-free	(.058)	.966	(.014)	.900	(0)	0	(0)	.293
			Non-uniform	(.184)	.437	(.026)	.323	(0)	.003	(0)	.051
			Uniform	(.070)	.934	(.018)	.832	(0)	0	(0)	.242
	12 items	100:900	DIF-free	(.187)	.733	(.184)	.718	(.381)	.667	(.434)	.853
			Non-uniform	(.477)	.014	(.477)	0	(.649)	0	(.829)	0
			Uniform	(.661)	.817	(.661)	.817	(.679)	.640	(.865)	.892
		250:750	DIF-free	(.026)	.149	(.004)	.057	(0)	0	(0)	0
			Non-uniform	(.024)	.732	(.002)	.709	(0)	.598	(0)	.795
			Uniform	(.028)	.088	(.004)	.012	(0)	.004	(0)	.004
		500:500	DIF-free	(.084)	.228	(.026)	.076	(0)	.003	(.004)	.009
			Non-uniform	(.060)	.114	(.016)	.024	(0)	0	(0)	.004
			Uniform	(.106)	.174	(.032)	.030	(0)	0	(.007)	.006

Power of Using Fixed-Factor Scaling Method for Detecting Non-Uniform DIF in the Equal Latent Trait Mean Condition

(Continued)

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Polytomous	6 items	100:900	DIF-free	(.071)	.145	(.018)	.040	(0)	0	(.006)	0
			Non-uniform	(.491)	.114	(.432)	.057	(.322)	0	(.559)	.092
			Uniform	(.073)	.142	(.020)	.036	(0)	0	(.006)	0
		250:750	DIF-free	(.058)	.386	(.014)	.142	(0)	0	(.006)	0
			Non-uniform	(.108)	.170	(.022)	.088	(0)	.004	(0)	.078
			Uniform	(.062)	.362	(.014)	.130	(0)	0	(0)	0
		500:500	DIF-free	(.092)	.416	(.014)	.184	(0)	0	(0)	0
			Non-uniform	(.170)	.146	(.056)	.047	(0)	.005	(0)	.050
			Uniform	(.094)	.404	(.018)	.168	(0)	0	(0)	0
	12 items	100:900	DIF-free	(.085)	.503	(.016)	.415	(0)	.253	(.005)	.480
			Non-uniform	(.293)	.262	(.246)	.147	(.249)	.115	(.438)	.137
			Uniform	(.085)	.247	(.016)	.084	(0)	0	(0)	.027
		250:750	DIF-free	(.088)	.418	(.014)	.140	(0)	.005	(0)	.058
			Non-uniform	(.064)	.298	(.004)	.074	(0)	0	(0)	0
			Uniform	(.094)	.380	(.014)	.108	(0)	0	(0)	.008
		500:500	DIF-free	(.146)	.564	(.010)	.206	(0)	.013	(0)	.013
			Non-uniform	(.106)	.494	(.002)	.158	(0)	0	(0)	0
			Uniform	(.152)	.582	(.012)	.218	(0)	.002	(0)	.011

Power of Using Fixed-Factor Scaling Method for Detecting Uniform DIF in the Equal Latent Trait Mean Condition

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Dichotomous	6 items	100:900	DIF-free	(.942)	.991	(.865)	.972	(.468)	.584	(.672)	.953
			Non-uniform	(.880)	.991	(.533)	.991	(0)	.564	(.002)	.945
			Uniform	(.882)	.990	(.539)	.990	(0)	.704	(.005)	.955
		250:750	DIF-free	(1)	1	(1)	1	(0)	.626	(.774)	1
			Non-uniform	(1)	1	(1)	1	(0)	.628	(.756)	1
			Uniform	(1)	1	(1)	1	(0)	.593	(.772)	1
		500:500	DIF-free	(1)	1	(1)	1	(0)	1	(.980)	1
			Non-uniform	(1)	1	(1)	1	(0)	1	(.980)	1
			Uniform	(1)	1	(1)	1	(0)	1	(.988)	1
	12 items	100:900	DIF-free	(.864)	1	(.374)	1	(0)	.673	(.002)	.867
			Non-uniform	(.967)	.995	(.914)	.995	(.592)	.527	(.817)	.745
			Uniform	(.865)	1	(.365)	1	(0)	.169	(0)	.346
		250:750	DIF-free	(1)	1	(1)	1	(0)	0	(.002)	.998
			Non-uniform	(1)	1	(1)	1	(0)	0	(0)	1
			Uniform	(1)	.997	(1)	.994	(0)	.601	(.002)	.971
		500:500	DIF-free	(1)	1	(1)	1	(0)	0	(.014)	1
			Non-uniform	(1)	1	(1)	1	(0)	0	(.014)	1
			Uniform	(1)	1	(1)	1	(0)	0	(.018)	1

Power of Using Fixed-Factor Scaling Method for Detecting Uniform DIF in the Equal Latent Trait Mean Condition

(Continued)

				Criterion value					
				LR test				ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	(.970)	1	(.727)	1	(0) 0	(0) 1
			Non-uniform	(.968)	1	(.711)	1	(0) 0	(.004) 1
			Uniform	(.970)	1	(.721)	1	(0) 0	(0) 1
		250:750	DIF-free	(1)	1	(1)	1	(0) 1	(.870) 1
			Non-uniform	(1)	1	(1)	1	(0) 1	(.862) 1
			Uniform	(1)	1	(1)	1	(0) 1	(.872) 1
		500:500	DIF-free	(1)	1	(1)	1	(0) 1	(1) 1
			Non-uniform	(1)	1	(1)	1	(0) 1	(1) 1
			Uniform	(1)	1	(1)	1	(0) 1	(1) 1
	12 items	100:900	DIF-free	(.968)	1	(.502)	1	(0) 0	(0) .700
			Non-uniform	(.965)	1	(.437)	.998	(0) 0	(0) .589
			Uniform	(.968)	1	(.501)	1	(0) 0	(0) .733
		250:750	DIF-free	(1)	1	(1)	1	(0) 0	(0) 1
			Non-uniform	(1)	1	(1)	1	(0) 0	(.004) 1
			Uniform	(1)	1	(1)	1	(0) 0	(0) 1
		500:500	DIF-free	(1)	1	(1)	1	(0) 0	(.002) 1
			Non-uniform	(1)	1	(1)	1	(0) 0	(.002) 1
			Uniform	(1)	1	(1)	1	(0) 0	(.002) 1

Power of Using Effects-Coded Scaling Method for Detecting Non-Uniform DIF in the Equal Latent Trait Mean Condition

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Dichotomous	6 items	100:900	DIF-free	(.862)	.940	(.848)	.898	(.633)	.632	(.824)	.835
			Non-uniform	(.799)	.945	(.766)	.910	(.552)	.698	(.778)	.885
			Uniform	(.950)	.922	(.939)	.922	(.635)	.675	(.856)	.907
		250:750	DIF-free	(.599)	.942	(.478)	.879	(.235)	.032	(.368)	.338
			Non-uniform	(.435)	.659	(.249)	.584	(.003)	.008	(.014)	.226
			Uniform	(.982)	.942	(.974)	.911	(.703)	.617	(.941)	.848
		500:500	DIF-free	(.490)	.992	(.313)	.968	(0)	0	(.049)	.733
			Non-uniform	(.454)	.469	(.276)	.381	(0)	.024	(.003)	.285
			Uniform	(.966)	.796	(.912)	.612	(0)	0	(.549)	.187
	12 items	100:900	DIF-free	(.835)	1	(.807)	.989	(.570)	.698	(.814)	.906
			Non-uniform	(.673)	.684	(.604)	.627	(.526)	.493	(.664)	.647
			Uniform	(.968)	.943	(.941)	.931	(.691)	.593	(.903)	.836
		250:750	DIF-free	(.425)	.972	(.241)	.947	(.075)	.486	(.101)	.684
			Non-uniform	(.246)	.766	(.117)	.704	(.033)	.366	(.041)	.523
			Uniform	(.969)	1	(.948)	.996	(.601)	.653	(.814)	.923
		500:500	DIF-free	(.432)	.921	(.234)	.851	(0)	.002	(.011)	.151
			Non-uniform	(.307)	.543	(.138)	.319	(0)	0	(.003)	.013
			Uniform	(.702)	.961	(.516)	.908	(0)	.002	(.019)	.288

Power of Using Effects-Coded Scaling Method for Detecting Non-Uniform DIF in the Equal Latent Trait Mean Condition

(Continued)

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Polytomous	6 items	100:900	DIF-free	(.251)	.387	(.125)	.146	(0)	0	(.007)	0
			Non-uniform	(.219)	.213	(.103)	.081	(0)	0	(.004)	.018
			Uniform	(.284)	.301	(.151)	.110	(0)	0	(.007)	0
		250:750	DIF-free	(.255)	.742	(.118)	.534	(0)	0	(.007)	.021
			Non-uniform	(.282)	.387	(.148)	.245	(0)	.006	(0)	.096
			Uniform	(.326)	.660	(.174)	.414	(0)	0	(.009)	.009
		500:500	DIF-free	(.298)	.848	(.144)	.700	(0)	0	(.007)	.089
			Non-uniform	(.316)	.417	(.140)	.287	(0)	.003	(0)	.047
			Uniform	(.398)	.762	(.230)	.574	(0)	0	(.008)	.051
	12 items	100:900	DIF-free	(.251)	.829	(.095)	.739	(0)	.248	(.004)	.404
			Non-uniform	(.130)	.634	(.035)	.472	(0)	.091	(0)	.115
			Uniform	(.273)	.745	(.089)	.622	(0)	.002	(0)	.054
		250:750	DIF-free	(.290)	.909	(.124)	.830	(0)	.107	(0)	.253
			Non-uniform	(.146)	.658	(.034)	.488	(0)	0	(0)	.007
			Uniform	(.354)	.904	(.142)	.816	(0)	0	(0)	.148
		500:500	DIF-free	(.360)	.972	(.138)	.922	(0)	.002	(0)	.221
			Non-uniform	(.182)	.822	(.030)	.618	(0)	0	(0)	.013
			Uniform	(.426)	.972	(.184)	.950	(0)	.006	(0)	.278

Power of Using Effects-Coded Scaling Method for Detecting Uniform DIF in the Equal Latent Trait Mean Condition

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–.01		–.002	
Dichotomous	6 items	100:900	DIF-free	(.932)	.992	(.915)	.962	(.632)	.625	(.896)	.890
			Non-uniform	(.893)	.962	(.888)	.873	(.580)	.490	(.865)	.675
			Uniform	(.893)	.833	(.857)	.809	(.538)	.692	(.829)	.892
		250:750	DIF-free	(.320)	.917	(.124)	.834	(0)	.333	(.003)	.523
			Non-uniform	(.470)	.951	(.232)	.879	(0)	.335	(.012)	.598
			Uniform	(.426)	.503	(.212)	.453	(0)	.399	(.015)	.588
		500:500	DIF-free	(.354)	.946	(.138)	.862	(0)	0	(.005)	.240
			Non-uniform	(.510)	.976	(.248)	.900	(0)	0	(.005)	.333
			Uniform	(.686)	.151	(.422)	.054	(0)	.004	(.037)	.009
	12 items	100:900	DIF-free	(.896)	.675	(.884)	.632	(.654)	.531	(.880)	.738
			Non-uniform	(.736)	.757	(.727)	.723	(.586)	.581	(.833)	.752
			Uniform	(.840)	.856	(.790)	.813	(.552)	.550	(.768)	.713
		250:750	DIF-free	(.103)	.504	(.016)	.406	(.006)	.298	(.011)	.419
			Non-uniform	(.333)	.729	(.269)	.699	(.272)	.566	(.416)	.728
			Uniform	(.349)	.989	(.157)	.979	(0)	.704	(.003)	.896
		500:500	DIF-free	(.080)	.228	(.008)	.080	(0)	.004	(0)	.004
			Non-uniform	(.110)	.514	(.037)	.407	(.021)	.349	(.051)	.447
			Uniform	(.508)	.731	(.212)	.555	(0)	0	(0)	.026

Power of Using Effects-Coded Scaling Method for Detecting Uniform DIF in the Equal Latent Trait Mean Condition

(Continued)

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Polytomous	6 items	100:900	DIF-free	(.261)	.941	(.106)	.830	(0)	0	(0)	.148
			Non-uniform	(.232)	.915	(.098)	.784	(0)	0	(.003)	.092
			Uniform	(.725)	.071	(.499)	.012	(0)	0	(.032)	0
		250:750	DIF-free	(.668)	1	(.416)	1	(0)	0	(.008)	.706
			Non-uniform	(.698)	1	(.440)	.998	(0)	0	(.015)	.752
			Uniform	(.926)	.046	(.804)	.010	(0)	0	(.065)	0
		500:500	DIF-free	(.894)	1	(.686)	1	(0)	0	(.014)	.970
			Non-uniform	(.892)	1	(.696)	1	(0)	0	(.022)	.976
			Uniform	(.982)	.038	(.896)	.010	(0)	0	(.182)	0
	12 items	100:900	DIF-free	(.074)	.105	(.016)	.018	(0)	0	(0)	0
			Non-uniform	(.065)	.091	(.014)	.012	(0)	0	(0)	0
			Uniform	(.401)	.336	(.145)	.156	(0)	0	(0)	0
		250:750	DIF-free	(.126)	.212	(.024)	.052	(0)	0	(0)	0
			Non-uniform	(.132)	.258	(.022)	.064	(0)	0	(0)	0
			Uniform	(.492)	.400	(.206)	.178	(0)	0	(0)	0
		500:500	DIF-free	(.174)	.310	(.024)	.100	(0)	0	(0)	0
			Non-uniform	(.214)	.460	(.034)	.154	(0)	0	(0)	0
			Uniform	(.574)	.444	(.224)	.166	(0)	0	(0)	0

Power of Using Marker-Variable Scaling Method for Detecting Non-Uniform DIF in the Unequal Latent Trait Mean Condition

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–.01		–.002	
Dichotomous	6 items	100:900	DIF-free	(.876)	.198	(.811)	.107	(.423)	0	(.660)	.016
			Non-uniform	(.314)	.917	(.207)	.830	(.120)	.456	(.176)	.596
			Uniform	(.997)	.955	(.997)	.882	(.576)	.009	(.987)	.482
		250:750	DIF-free	(.803)	.168	(.726)	.079	(0)	0	(.376)	0
			Non-uniform	(.412)	.977	(.239)	.891	(0)	.010	(.028)	.225
			Uniform	(1)	1	(1)	1	(.980)	.036	(1)	.986
		500:500	DIF-free	(.936)	.132	(.892)	.034	(0)	0	(.572)	.005
			Non-uniform	(.534)	.998	(.280)	.996	(0)	.020	(0)	.422
			Uniform	(1)	1	(1)	1	(1)	.928	(1)	1
	12 items	100:900	DIF-free	(.764)	.922	(.648)	.909	(.012)	.644	(.021)	.849
			Non-uniform	(.936)	.991	(.904)	.982	(.556)	.687	(.708)	.908
			Uniform	(1)	1	(.995)	.995	(.647)	.660	(.956)	.919
		250:750	DIF-free	(.868)	.535	(.771)	.453	(0)	.029	(.112)	.540
			Non-uniform	(.841)	.484	(.678)	.246	(0)	0	(.028)	0
			Uniform	(1)	.997	(1)	.994	(.361)	.777	(.998)	.960
		500:500	DIF-free	(.980)	.636	(.910)	.588	(0)	.085	(.223)	.698
			Non-uniform	(.932)	.488	(.800)	.180	(0)	0	(.004)	0
			Uniform	(1)	1	(1)	.998	(.895)	.804	(.998)	.993

Power of Using Marker-Variable Scaling Method for Detecting Non-Uniform DIF in the Unequal Latent Trait Mean Condition (Continued)

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Polytomous	6 items	100:900	DIF-free	(.252)	.091	(.083)	.032	(0)	0	(.006)	0
			Non-uniform	(.072)	.235	(.020)	.078	(0)	0	(0)	0
			Uniform	(.924)	.235	(.829)	.090	(0)	0	(.098)	0
		250:750	DIF-free	(.386)	.192	(.156)	.068	(0)	0	(0)	0
			Non-uniform	(.046)	.492	(.008)	.210	(0)	0	(0)	0
			Uniform	(.994)	.248	(.972)	.100	(0)	0	(.418)	0
		500:500	DIF-free	(.516)	.244	(.240)	.098	(0)	0	(0)	0
			Non-uniform	(.044)	.765	(.012)	.386	(0)	0	(0)	.006
			Uniform	(1)	.522	(1)	.268	(0)	0	(.862)	.002
	12 items	100:900	DIF-free	(.373)	.983	(.114)	.949	(0)	.454	(0)	.715
			Non-uniform	(.683)	.144	(.556)	.033	(.308)	0	(.493)	0
			Uniform	(.970)	.963	(.890)	.943	(0)	.307	(0)	.479
		250:750	DIF-free	(.588)	.824	(.240)	.790	(0)	.005	(0)	.055
			Non-uniform	(.432)	.138	(.160)	.020	(0)	0	(0)	0
			Uniform	(.998)	.954	(.996)	.914	(0)	.004	(.006)	.632
		500:500	DIF-free	(.812)	.932	(.452)	.926	(0)	.004	(0)	.393
			Non-uniform	(.526)	.090	(.190)	.004	(0)	0	(0)	0
			Uniform	(1)	1	(1)	.990	(0)	.012	(.196)	.966

Power of Using Marker-Variable Scaling Method for Detecting Uniform DIF in the Unequal Latent Trait Mean Condition

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Dichotomous	6 items	100:900	DIF-free	(.013)	.547	(.003)	.453	(0)	.407	(0)	.556
			Non-uniform	(.889)	1	(.648)	1	(0)	.254	(.032)	.535
			Uniform	(.971)	.860	(.965)	.740	(.593)	.308	(.901)	.423
		250:750	DIF-free	(.002)	.863	(0)	.808	(0)	.602	(0)	.766
			Non-uniform	(.993)	.996	(.959)	.989	(.059)	.206	(.355)	.870
			Uniform	(.992)	.952	(.983)	.886	(.347)	.416	(.969)	.609
		500:500	DIF-free	(.024)	.724	(0)	.348	(0)	.002	(0)	.011
			Non-uniform	(1)	.997	(1)	.980	(0)	.582	(.754)	.969
			Uniform	(1)	.924	(1)	.789	(.726)	.004	(1)	.276
	12 items	100:900	DIF-free	(.015)	.783	(.004)	.652	(0)	.486	(0)	.657
			Non-uniform	(1)	1	(.940)	1	(.508)	.708	(.659)	.923
			Uniform	(1)	.500	(.967)	.188	(.395)	0	(.497)	0
		250:750	DIF-free	(.101)	.516	(.101)	.143	(.255)	0	(.324)	.004
			Non-uniform	(.996)	1	(.983)	1	(0)	.547	(.011)	.688
			Uniform	(1)	.945	(.998)	.830	(0)	.357	(.614)	.439
		500:500	DIF-free	(.070)	.823	(.002)	.394	(0)	0	(0)	0
			Non-uniform	(1)	1	(1)	1	(.014)	.019	(.407)	.529
			Uniform	(1)	.776	(1)	.592	(.006)	0	(.988)	.005

Power of Using Marker-Variable Scaling Method for Detecting Uniform DIF in the Unequal Latent Trait Mean Condition

(Continued)

				Criterion value					
				LR test				ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	(.500)	1	(.163)	1	(0) 0	(0) 1
			Non-uniform	(1)	1	(1)	1	(0) .673	(.990) 1
			Uniform	(1)	.744	(1)	.414	(0) 0	(1) .002
		250:750	DIF-free	(.812)	1	(.446)	1	(0) .102	(0) 1
			Non-uniform	(1)	1	(1)	1	(.326) 1	(1) 1
			Uniform	(1)	.964	(1)	.836	(.848) 0	(1) .032
		500:500	DIF-free	(.886)	1	(.604)	1	(0) .794	(0) 1
			Non-uniform	(1)	1	(1)	1	(.774) 1	(1) 1
			Uniform	(1)	.998	(1)	.992	(1) 0	(1) .314
	12 items	100:900	DIF-free	(.534)	.998	(.108)	.998	(0) 0	(0) .289
			Non-uniform	(1)	1	(1)	1	(.116) 0	(.495) .998
			Uniform	(1)	.707	(1)	.230	(0) 0	(.655) 0
		250:750	DIF-free	(.840)	1	(.338)	1	(0) 0	(0) .998
			Non-uniform	(1)	1	(1)	1	(0) .728	(1) 1
			Uniform	(1)	.948	(1)	.686	(0) 0	(1) 0
		500:500	DIF-free	(.888)	1	(.468)	1	(0) 0	(0) 1
			Non-uniform	(1)	1	(1)	1	(0) .998	(1) 1
			Uniform	(1)	.998	(1)	.948	(.008) 0	(1) 0

Power of Using Fixed-Factor Scaling Method for Detecting Non-Uniform DIF in the Unequal Latent Trait Mean Condition

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Dichotomous	6 items	100:900	DIF-free	(.868)	.902	(.854)	.713	(.597)	0	(.859)	.030
			Non-uniform	(.817)	.865	(.784)	.798	(.593)	.563	(.809)	.707
			Uniform	(.924)	.868	(.919)	.658	(.677)	0	(.896)	.028
		250:750	DIF-free	(.730)	1	(.490)	1	(.002)	0	(.039)	.749
			Non-uniform	(.528)	.874	(.279)	.768	(0)	.004	(.011)	.322
			Uniform	(.682)	1	(.424)	.992	(0)	0	(.017)	.695
		500:500	DIF-free	(.748)	1	(.506)	1	(0)	0	(.038)	.950
			Non-uniform	(.536)	.929	(.324)	.857	(0)	.006	(.002)	.521
			Uniform	(.696)	1	(.448)	.998	(0)	0	(.017)	.924
	12 items	100:900	DIF-free	(.454)	.553	(.336)	.470	(.293)	.343	(.395)	.461
			Non-uniform	(.985)	.982	(.970)	.964	(.686)	.667	(.924)	.896
			Uniform	(.811)	.702	(.769)	.675	(.598)	.605	(.809)	.772
		250:750	DIF-free	(.744)	.459	(.352)	.436	(0)	0	(0)	.012
			Non-uniform	(.731)	.066	(.322)	.018	(0)	0	(0)	0
			Uniform	(.690)	.200	(.290)	.173	(0)	.006	(.002)	.019
		500:500	DIF-free	(.722)	.395	(.394)	.372	(0)	0	(0)	.026
			Non-uniform	(.736)	.062	(.402)	.012	(0)	0	(0)	0
			Uniform	(.682)	.151	(.338)	.139	(0)	.006	(0)	.023

Power of Using Fixed-Factor Scaling Method for Detecting Non-Uniform DIF in the Unequal Latent Trait Mean Condition

(Continued)

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–.01		–.002	
Polytomous	6 items	100:900	DIF-free	(.034)	.157	(.006)	.056	(0)	0	(.007)	0
			Non-uniform	(.052)	.109	(.012)	.036	(0)	0	(.006)	.024
			Uniform	(.038)	.149	(.008)	.054	(0)	0	(.020)	0
		250:750	DIF-free	(.042)	.426	(0)	.180	(0)	0	(0)	.002
			Non-uniform	(.064)	.143	(.012)	.041	(0)	0	(0)	.005
			Uniform	(.044)	.410	(0)	.166	(0)	0	(0)	.003
		500:500	DIF-free	(.052)	.508	(.012)	.226	(0)	0	(0)	.002
			Non-uniform	(.134)	.124	(.026)	.042	(0)	0	(0)	.014
			Uniform	(.054)	.478	(.018)	.212	(0)	0	(0)	.002
	12 items	100:900	DIF-free	(.040)	.221	(0)	.054	(0)	0	(0)	.007
			Non-uniform	(.482)	.154	(.447)	.024	(.432)	0	(.675)	0
			Uniform	(.044)	.464	(.002)	.341	(0)	.259	(0)	.371
		250:750	DIF-free	(.048)	.285	(0)	.082	(0)	0	(0)	.023
			Non-uniform	(.040)	.210	(0)	.034	(0)	0	(0)	0
			Uniform	(.048)	.262	(0)	.058	(0)	0	(0)	.003
		500:500	DIF-free	(.092)	.441	(.010)	.138	(0)	0	(0)	0
			Non-uniform	(.068)	.370	(.006)	.098	(0)	0	(0)	0
			Uniform	(.094)	.451	(.010)	.136	(0)	0	(0)	0

Power of Using Fixed-Factor Scaling Method for Detecting Uniform DIF in the Unequal Latent Trait Mean Condition

				Criterion value					
				LR test				ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Dichotomous	6 items	100:900	DIF-free	(1)	1	(.996)	1	(.095)	.560	(.989)	1
			Non-uniform	(.985)	1	(.964)	1	(.694)	.500	(.948)	.986
			Uniform	(1)	1	(1)	1	(.010)	.617	(1)	.983
		250:750	DIF-free	(1)	.996	(1)	.993	(1)	.890	(1)	.988
			Non-uniform	(1)	.989	(1)	.985	(.887)	.927	(.982)	.996
			Uniform	(1)	.996	(1)	.996	(1)	.902	(1)	.984
		500:500	DIF-free	(1)	1	(1)	1	(1)	1	(1)	1
			Non-uniform	(1)	1	(1)	1	(1)	1	(1)	1
			Uniform	(1)	1	(1)	1	(1)	1	(1)	1
	12 items	100:900	DIF-free	(1)	1	(1)	1	(.356)	.333	(.915)	.970
			Non-uniform	(1)	1	(1)	1	(.644)	.667	(.931)	1
			Uniform	(1)	1	(.994)	1	(.562)	.129	(.922)	.968
		250:750	DIF-free	(1)	1	(1)	1	(.002)	.517	(1)	.976
			Non-uniform	(1)	1	(1)	1	(0)	.242	(.998)	1
			Uniform	(1)	1	(1)	.995	(0)	.425	(1)	.995
		500:500	DIF-free	(1)	.997	(1v	.994	(.890)	.865	(1)	.974
			Non-uniform	(.998)	.997	(.998)	.991	(.808)	.856	(.998)	.965
			Uniform	(1)	1	(1)	1	(.920)	.998	(1)	1

Power of Using Fixed-Factor Scaling Method for Detecting Uniform DIF in the Unequal Latent Trait Mean Condition

(Continued)

				Criterion value					
				LR test				ΔCFI test	
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01	–0.002
Polytomous	6 items	100:900	DIF-free	(1)	1	(1)	1	(.506) 1	(1) 1
			Non-uniform	(1)	.994	(1)	.994	(.432) .882	(1) .983
			Uniform	(1)	1	(1)	1	(.478) 1	(1) 1
		250:750	DIF-free	(1)	1	(1)	1	(1) 1	(1) 1
			Non-uniform	(1)	1	(1)	1	(1) 1	(1) 1
			Uniform	(1)	1	(1)	1	(1) 1	(1) 1
		500:500	DIF-free	(1)	1	(1)	1	(1) 1	(1) 1
			Non-uniform	(1)	1	(1)	1	(1) 1	(1) 1
			Uniform	(1)	1	(1)	1	(1) 1	(1) 1
	12 items	100:900	DIF-free	(1)	1	(1)	.998	(0) .189	(1) .981
			Non-uniform	(1)	1	(1)	1	(.002) .006	(.996) .998
			Uniform	(1)	1	(1)	1	(0) .006	(1) 1
		250:750	DIF-free	(1)	1	(1)	1	(.124) 1	(1) 1
			Non-uniform	(1)	1	(1)	1	(.058) 1	(1) 1
			Uniform	(1)	1	(1)	1	(.120) 1	(1) 1
		500:500	DIF-free	(1)	1	(1)	1	(1) 1	(1) 1
			Non-uniform	(1)	1	(1)	1	(1) 1	(1) 1
			Uniform	(1)	1	(1)	1	(1) 1	(1) 1

Power of Using Effects-Coded Scaling Method for Detecting Non-Uniform DIF in the Equal Latent Trait Mean Condition

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Dichotomous	6 items	100:900	DIF-free	(.941)	.919	(.926)	.900	(.644)	.649	(.853)	.919
			Non-uniform	(.952)	.994	(.940)	.982	(.621)	.644	(.845)	.911
			Uniform	(.970)	.977	(.964)	.977	(.691)	.623	(.926)	.892
		250:750	DIF-free	(.643)	.802	(.529)	.676	(0)	.340	(.283)	.522
			Non-uniform	(.946)	.721	(.923)	.585	(.596)	.267	(.812)	.405
			Uniform	(.976)	.948	(.950)	.917	(.448)	.622	(.887)	.902
		500:500	DIF-free	(.771)	.790	(.656)	.626	(0)	0	(.304)	.113
			Non-uniform	(.608)	.634	(.457)	.466	(0)	.008	(.061)	.110
			Uniform	(.991)	.323	(.975)	.241	(.592)	.247	(.965)	.326
	12 items	100:900	DIF-free	(.838)	1	(.760)	1	(.526)	.659	(.727)	.909
			Non-uniform	(.977)	.984	(.977)	.974	(.697)	.636	(.928)	.895
			Uniform	(1)	.983	(1)	.983	(.671)	.711	(.890)	.950
		250:750	DIF-free	(1)	1	(.991)	.985	(.695)	.652	(.964)	.928
			Non-uniform	(.329)	.825	(.162)	.694	(.046)	.004	(.060)	.160
			Uniform	(.970)	.918	(.932)	.892	(.598)	.533	(.791)	.855
		500:500	DIF-free	(.540)	.836	(.366)	.728	(0)	.007	(.010)	.440
			Non-uniform	(.330)	.846	(.152)	.732	(0)	0	(.003)	.135
			Uniform	(.887)	.987	(.768)	.977	(.006)	.734	(.092)	.926

Power of Using Effects-Coded Scaling Method for Detecting Non-Uniform DIF in the Unequal Latent Trait Mean Condition

(Continued)

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Polytomous	6 items	100:900	DIF-free	(.231)	.307	(.122)	.127	(0)	0	(.011)	0
			Non-uniform	(.247)	.205	(.118)	.093	(0)	0	(0)	.011
			Uniform	(.469)	.161	(.303)	.048	(0)	0	(.038)	0
		250:750	DIF-free	(.238)	.716	(.106)	.470	(0)	0	(.003)	.015
			Non-uniform	(.275)	.359	(.116)	.213	(0)	0	(0)	.024
			Uniform	(.578)	.428	(.404)	.222	(0)	0	(.030)	0
		500:500	DIF-free	(.332)	.812	(.164)	.620	(0)	0	(.009)	.081
			Non-uniform	(.352)	.394	(.178)	.244	(0)	0	(.006)	.027
			Uniform	(.778)	.510	(.562)	.302	(0)	0	(.053)	.012
	12 items	100:900	DIF-free	(.233)	.795	(.098)	.642	(0)	0	(0)	.030
			Non-uniform	(.195)	.593	(.087)	.391	(.078)	0	(.141)	.005
			Uniform	(.369)	.821	(.171)	.730	(0)	0	(0)	.043
		250:750	DIF-free	(.274)	.856	(.088)	.734	(0)	0	(0)	.062
			Non-uniform	(.122)	.648	(.022)	.440	(0)	0	(0)	.002
			Uniform	(.426)	.892	(.204)	.814	(0)	0	(0)	.121
		500:500	DIF-free	(.376)	.948	(.166)	.865	(0)	0	(0)	.180
			Non-uniform	(.182)	.770	(.044)	.628	(0)	0	(0)	.004
			Uniform	(.574)	.972	(.314)	.946	(0)	0	(0)	.298

Power of Using Effects-Coded Scaling Method for Detecting Uniform DIF in the Unequal Latent Trait Mean Condition

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Dichotomous	6 items	100:900	DIF-free	(.960)	.926	(.929)	.852	(.563)	.607	(.830)	.821
			Non-uniform	(.962)	1	(.962)	.923	(.680)	.519	(.932)	.852
			Uniform	(.795)	.815	(.761)	.815	(.689)	.750	(.878)	.917
		250:750	DIF-free	(.992)	.986	(.984)	.971	(.717)	.789	(.945)	.945
			Non-uniform	(.996)	1	(.987)	1	(.757)	.806	(.949)	.919
			Uniform	(.946)	.908	(.938)	.885	(.619)	.746	(.888)	.885
		500:500	DIF-free	(.875)	.981	(.763)	.969	(.002)	.836	(.269)	.975
			Non-uniform	(.994)	.996	(.982)	.965	(0)	.782	(.770)	.953
			Uniform	(.706)	1	(.650)	.995	(.545)	.890	(.709)	.987
	12 items	100:900	DIF-free	(.928)	.800	(.895)	.760	(.615)	.571	(.885)	.714
			Non-uniform	(.751)	1	(.709)	1	(.516)	.759	(.665)	1
			Uniform	(1)	.667	(1)	.667	(.675)	.500	(.906)	.500
		250:750	DIF-free	(.975)	.921	(.953)	.890	(.715)	.574	(.911)	.798
			Non-uniform	(.595)	.995	(.357)	.995	(.135)	.761	(.168)	.955
			Uniform	(.786)	.950	(.746)	.950	(.591)	.736	(.740)	.925
		500:500	DIF-free	(.622)	1	(.390)	1	(0)	.787	(0)	.984
			Non-uniform	(.908)	.989	(.801)	.989	(.011)	.779	(.063)	.929
			Uniform	(.334)	1	(.150)	1	(0)	.833	(.003)	1

Power of Using Effects-Coded Scaling Method for Detecting Uniform DIF in the Unequal Latent Trait Mean Condition

(Continue)

				Criterion value							
				LR test				ΔCFI test			
Response	Scale size	Samples	Type of anchor	Uncorrected *p*		Corrected *p*		–0.01		–0.002	
Polytomous	6 items	100:900	DIF-free	(.910)	.996	(.749)	.978	(0)	0	(.045)	.540
			Non-uniform	(.998)	.984	(.990)	.970	(0)	.327	(.478)	.923
			Uniform	(.070)	.323	(.024)	.132	(0)	0	(0)	0
		250:750	DIF-free	(.994)	1	(.984)	1	(0)	0	(.396)	.982
			Non-uniform	(1)	1	(1)	1	(0)	.086	(.996)	1
			Uniform	(.092)	.540	(.024)	.316	(0)	0	(0)	.007
		500:500	DIF-free	(1)	1	(.994)	1	(0)	.018	(.738)	1
			Non-uniform	(1)	1	(1)	1	(.014)	.822	(1)	1
			Uniform	(.130)	.716	(.052)	.522	(0)	0	(0)	.026
	12 items	100:900	DIF-free	(.946)	.892	(.773)	.719	(0)	0	(0)	.004
			Non-uniform	(.968)	.963	(.858)	.834	(.108)	0	(.173)	.008
			Uniform	(.469)	.503	(.191)	.222	(0)	0	(0)	0
		250:750	DIF-free	(1)	.994	(.990)	.964	(0)	0	(.040)	.036
			Non-uniform	(1)	1	(1)	1	(0)	0	(.144)	.200
			Uniform	(.796)	.746	(.502)	.502	(0)	0	(0)	0
		500:500	DIF-free	(1)	1	(1)	1	(0)	0	(.170)	.234
			Non-uniform	(1)	1	(1)	1	(0)	0	(.778)	.780
			Uniform	(.902)	.860	(.702)	.666	(0)	0	(0)	0